THE UNIVERSITY *of* EDINBURGH

Models and metaphors in neuroscience :

The role of dopamine in reinforcement learning as a case study

Robert Kyle

Doctor of Philosophy
Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh
2011

Abstract

Neuroscience makes use of many metaphors in its attempt to explain the relationship between our brain and our behaviour. In this thesis I contrast the most commonly used metaphor - that of computation driven by neuron action potentials - with an alternative view which seeks to understand the brain in terms of an agent learning from the reward signalled by neuromodulators.

To explore this reinforcement learning model I construct computational models to assess one of its key claims — that the neurotransmitter dopamine signals unexpected reward, and that this signal is used by the brain to learn control of our movements and drive goal-directed behaviour.

In this thesis I develop a selection of computational models that are motivated by either theoretical concepts or experimental data relating to the effects of dopamine.

The first model implements a published dopamine-modulated spike timing-dependent plasticity mechanism but is unable to correctly solve the distal reward problem. I analyse why this model fails and suggest solutions.

The second model, more closely linked to the empirical data attempts to investigate the relative contributions of firing rate and synaptic conductances to synaptic plasticity. I use experimental data to estimate how model neurons will be affected by dopamine modulation, and use the resulting computational model to predict the effect of dopamine on synaptic plasticity. The results suggest that dopamine modulation of synaptic conductances is more significant than modulation of excitability.

The third model demonstrates how simple assumptions about the anatomy of the basal ganglia, and the electrophysiological effects of dopamine modulation can lead to reinforcement learning like behaviour. The model makes the novel prediction that working memory is an emergent feature of a reinforcement learning process.

In the course of producing these models I find that both theoretically and empirically based models suffer from methodological problems that make it difficult to adequately support such fundamental claims as the reinforcement learning hypothesis.

The conclusion that I draw from the modelling work is that it is neither possible, nor desirable to falsify the theoretical models used in neuroscience. Instead I argue that models and metaphors can be valued by how useful they are, independently of their truth.

As a result I suggest that we ought to encourage a plurality of models and metaphors in neuroscience.

In Chapter 7 I attempt to put this into practice by reviewing the other transmitter systems that modulate dopamine release, and use this as a basis for exploring the context of dopamine modulation and reward-driven behaviour. I draw on evidence to suggest that dopamine modulation can be seen as part of an extended stress response, and that the function of dopamine is to encourage the individual to engage in behaviours that take it away from homeostasis. I also propose that the function of dopamine can be interpreted in terms of behaviourally defining self and non-self, much in the same way as inflammation and antibody responses are said to do in immunology.

Acknowledgements

I'd like to thank David Willshaw, my supervisor, who has always been supportive and encouraging, and who gave me the freedom to follow my own interests. A big thank you to my external supervisor Daniel Durstewitz, who was willing to supervise my work at a distance, who offered me his invaluable neuroscientific expertise, and who pushed me to get those models working! I am grateful to have been a welcome visitor to his workplace in Plymouth and Mannheim. I would like to thank Gareth Leng who provided me with the constructive feedback and debate that helped shape chapter 7. Thank you also to Andrew Gillies who at the very start of my PhD encouraged me to follow the questions that interested me.

Thank you Jeff Mitchell and Chris Ball who were adept at pointing out holes in my reasoning, and were even willing to provide feedback on my thesis. Thanks to the many students I shared offices with, in Edinburgh, Plymouth, and Mannheim. Thanks to Cian O'Donnell, Matty Chalk, Hugh Pastoll, Tim O'Leary, Andy Fugard, and John Davey, who were all willing to discuss the more philosophical points of my thesis.

A special thanks go to the secretarial and support staff, particularly Pat Ferguson who made my transition across institutes and funding bodies as pain free as possible. I am grateful to the academic and secretarial staff for providing me with the opportunity to work in such a stimulating environment. Thank you to the ESRC, and EPSRC whose financial support made my studies and travels possible in the first place.

And thanks to my friends and family, who have supported me all the way.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Robert Kyle*)

Table of Contents

1 Introduction **1**

1.1 Précis . . . 1

1.2 Models and metaphors in neuroscience . . . 1

1.2.1 Introduction . . . 1

1.2.2 The *metaphor* : Computation . . . 2

1.2.3 The *model* : Computation using spikes . . . 2

1.2.4 A brief history of spike-based models in neuroscience . . . 3

1.2.5 Summary . . . 6

1.3 Beyond neural computation . . . 6

1.3.1 Beyond spike-based models . . . 7

1.3.2 The computational metaphor . . . 10

1.4 A separation of timescales . . . 11

1.5 Neuromodulation as a means for explaining behaviour . . . 16

1.5.1 A definition . . . 16

1.5.2 How do neuromodulators have a consistent effect on behaviour? 17

1.5.3 The effect of neuromodulation upon behaviour . . . 19

1.6 Using neuromodulation to explain decision making . . . 19

1.7 The questions in this thesis . . . 21

1.8 Thesis plan . . . 22

2 Dopamine and Reinforcement Learning: A literature review **24**

2.1 Novel contributions . . . 24

2.2 Introduction . . . 24

2.3 The dopaminergic reinforcement learning hypothesis . . . 24

2.4 Reinforcement learning theory . . . 27

2.4.1 The Temporal Differences algorithm . . . 27

2.4.2 Backpropagation of reward prediction error . . . 29

2.5 Existing biologically inspired models of reinforcement learning 30
2.6 Mechanisms underlying dopamine modulated reinforcement learning 30
2.7 Literature review 32
2.7.1 Dopamine anatomy 32
2.7.2 Dopamine release dynamics 36
2.7.3 Diffusion and re-uptake 36
2.7.4 Binding dynamics 38
2.7.5 Intracellular cascades 44
2.7.6 Electrophysiological effects 45
2.8 Summary 46

3 A dopamine-modulated STDP model of Reinforcement Learning 48
3.1 Novel contributions 48
3.2 Introduction 48
3.3 Literature review : Existing models of dopamine modulated reinforcement learning 48
3.4 Dopamine modulation of spike-timing dependent plasticity 49
3.5 Model 1 50
3.5.1 Neuron dynamics 50
3.5.2 Network and connectivity 50
3.5.3 Dopamine and reward 51
3.5.4 Synaptic plasticity 51
3.5.5 Results 52
3.6 Model 2 54
3.6.1 Phase 1 : The bell (US) 55
3.6.2 Phase 2 : The bell then the food (CS1 $\rightarrow$ US) 56
3.6.3 Phase 3 : Bell 2, then Bell 1, then the food (CS2 $\rightarrow$ CS1 $\rightarrow$ US) 56
3.7 The practicalities of extending upon model 2 56
3.8 Problems with the model 58
3.8.1 The US is not necessary for learning the CS 58
3.8.2 Learning is too quick 58
3.8.3 Learning is indiscriminate 61
3.8.4 US response is depressed by an unrealistic mechanism 61
3.8.5 Dopamine release is happening in an artificial way 64
3.8.6 Reward prediction error is not backpropagated 66

3.9 Potential Solutions . . . 66
3.9.1 Novelty . . . 66
3.9.2 Implementing backpropagation of reward prediction error . . . 67
3.10 Summary of findings . . . 68
3.11 Discussion . . . 69
3.11.1 Theoretical Issues . . . 69
3.11.2 Practical Issues . . . 72
3.12 Conclusion . . . 73

4 The effect of dopamine modulation on a model of synaptic plasticity 74
4.1 Novel contribution . . . 74
4.2 Empirical Modelling . . . 74
4.2.1 Investigating reinforcement learning with an empirical model 75
4.2.2 Setting up a model . . . 76
4.2.3 Signs of successful reinforcement learning . . . 76
4.3 Literature review . . . 77
4.3.1 The effect of dopamine on synaptic plasticity: Empirical work 78
4.3.2 The effect of dopamine on synaptic plasticity: Computational modelling . . . 79
4.3.3 Summary . . . 80
4.4 Methods . . . 80
4.4.1 The neuron model . . . 81
4.4.2 The plasticity model . . . 88
4.4.3 Simulated dopamine modulation . . . 89
4.4.4 The plasticity protocols . . . 95
4.5 Results . . . 98
4.6 Discussion . . . 102
4.6.1 The neuron model . . . 102
4.6.2 The plasticity model . . . 104
4.6.3 The methodology . . . 105
4.7 Conclusions . . . 106
4.8 A third approach . . . 106

5 Diffusion based Reinforcement Learning model 109
5.1 Novel contribution . . . 109
5.2 Introduction . . . 109

5.3 Aims of the model . . . 111
5.4 Literature review . . . 111
5.4.1 The neurophysiology of reinforcement learning . . . 111
5.4.2 Existing computational models . . . 112
5.4.3 The link between working memory and reinforcement learning 113
5.4.4 The neurophysiology of reinforcement learning : key observations . . . 113
5.4.5 The role of dopamine in a reinforcement learning circuit . . . 114
5.5 Constructing the model . . . 115
5.6 Learning working memory from reward . . . 116
5.6.1 Method . . . 116
5.6.2 Results . . . 118
5.7 Back propagation of predictions . . . 122
5.7.1 Method . . . 123
5.7.2 Results . . . 123
5.8 Back propagation of prediction error . . . 126
5.8.1 Method . . . 128
5.8.2 Results . . . 130
5.9 Discussion . . . 130
5.9.1 A review of the modelling approach . . . 134
5.10 Conclusion . . . 135

6 Discussion **136**
6.1 Review of the modelling work . . . 136
6.1.1 Model 1 : Reinforcement learning by dopamine modulated spike-timing dependent plasticity . . . 136
6.1.2 Model 2 : The effect of dopamine upon synaptic plasticity . . 136
6.1.3 Model 3 : A conceptual model of physiological basis of reinforcement learning . . . 137
6.2 Discussion . . . 137
6.2.1 Neuromodulation and reinforcement learning as a paradigm . 138
6.3 Support for the neuromodulation and reinforcement learning model . 139
6.4 Evidence against the neuromodulation and reinforcement learning model 141
6.4.1 Dopamine neurons fire phasically during sleep . . . 141
6.4.2 Dopamine is also released by noradrenaline neurons . . . 141

6.4.3 Not all behaviour can be understood in terms of reward 142
6.4.4 Dopamine may be too slow for reward learning 142
6.5 An alternative view of dopamine modulation 142
6.6 Is dopamine a unitary entity? 143
6.7 Summary 143

7 Putting dopamine and reward in context **145**
7.1 Novel contribution 145
7.2 Précis 145
7.3 Introduction 146
7.3.1 The biological case 147
7.3.2 The historical case 150
7.3.3 Summary 151
7.4 Dopamine modulation in context 151
7.5 Stress 151
7.5.1 Dopamine and stress : Positive and negative feedback loops . 154
7.5.2 Dopamine as an anticipatory stress system 155
7.5.3 Dopamine as a generator of allostasis 155
7.5.4 Summary 157
7.6 Immunity 157
7.6.1 Causal interactions between the immunity and reward 157
7.6.2 Self and non-self in Immunity and reward 159
7.6.3 Linking the Immune and Psychological selves 161
7.6.4 Dopamine and reward as a self(ish) circuit 161
7.6.5 Summary 162
7.7 Dopamine: reward, wanting, or agency? 163

8 Conclusion **165**
8.1 Summary of findings 165
8.2 How we can compare models 166
8.3 Relating brain processes to human psychology 166
8.4 Hiding the world in the brain 167

A Chapter 3 model parameters **168**

B Chapter 4 model code **170**

C Chapter 4 model parameters **185**

D Parameters used in preliminary dopamine diffusion and re-uptake model **188**

Bibliography **190**

Chapter 1

Introduction

1.1 Précis

In this thesis I examine the models and metaphors that are used in neuroscience. In particular I am interested in a recently introduced model — that of the brain as a reinforcement learning circuit. In this introductory chapter I attempt to motivate why this model is interesting, and how it differs from the models and metaphors we currently use.

1.2 Models and metaphors in neuroscience

1.2.1 Introduction

To explain the motivation behind my thesis topic, I first need to explain why I, or why neuroscientists as a whole choose to study the brain at all. Why is the brain so intrinsically interesting to scientists?

It is interesting because since Descartes it has been thought to be the seat of the mind — the physiological home of our identity, and place from where our thoughts, and decisions are initiated (for a review see (Lokhorst, 2009)). Our desire to understand the brain reflects our desire to understand ourselves, and why we do what we do. Many scientists choose to understand the relationship between mind and brain in different ways — some study perception, some the process of decision making, while others focus on learning and memory. What unites these different approaches, is that *they all attempt to use processes in the brain to explain human behaviour and our first-person experience*. In fact Eric Kandel, a nobel prize-winning neuroscientist defines

neuroscience as *the scientific attempt to explain the relationship between the brain and our behaviour* (Kandel et al., 2000).

If our aim as neuroscientists is to explain this relationship, then I will start out by looking at how this is usually done in neuroscience.

1.2.2 The *metaphor* : Computation

Historically there have been many different metaphors used to explain how processes in our brain relate to our behaviour, but the most common metaphor in use today is the computational metaphor.

The use of metaphors is more common amongst theoretical neuroscientists than experimentalists or clinicians, but these metaphors are important as they effect the explanations that we create, the experiments we design, and the clinical interventions that we attempt.

The computational metaphor compares processes in the brain to those of a computer — *information* arrives from the periphery as *input*, undergoes some process of *computation* in our brain, and is *output* as motor signals to our muscles. In this framework our role as neuroscientists is to examine the brain in much the same way as a computer scientist would reverse-engineer a computer. If we control the input to the brain and observe its output, then we can try to deconstruct the internal processes and determine what caused the ouput that we observe.

1.2.3 The *model* : Computation using spikes

When we look at the brain one of the most significant things that we observe are the millions of action potentials that occur every second. In most cases action potentials are all-or-nothing phenomena, and it is often assumed that these binary signals are the mechanism by which our brain transforms input into output. The presence of these action potentials is often inferred as evidence of computation. Like the binary signals in a computer, spikes are though to form the basis of a digital computation that happens between input (our senses) and output (our actions).

One example of how this metaphor of computation is applied to spiking neurons is the neural coding hypothesis — a claim that the action potentials and spike rates of neurons in the brain represent our internal states during the process of computation (for examples of the neural coding hypothesis see (Dayan and Abbott, 2001)). The implication here is that neuroscience should proceed by observing the patterns of spikes in

neurons, and decode these patterns to find the algorithms that generate our behaviour.

This view is widespread in neuroscience, particularly among theorists, and it is not uncommon to hear experimentalists inadvertently talk about their results in terms of information transfer, encoding, and storage. In fact this hypothesis is so well accepted in the theoretical community that theoretical neuroscience and computational neuroscience are often treated as synonymous.

But where does this view come from? Is it reasonable to try to explain all of our behaviour in terms of neural codes?

1.2.4 A brief history of spike-based models in neuroscience

As is often the case in science, the models and metaphors we use in neuroscience relate to the kind of measurements we have historically been able to make. The belief that the neuron is the most significant unit in the nervous system stems in part from the observations of Luigi Galvani in 1786, who discovered that electrical charge could cause the movement of a pair of frogs legs (for a review, see (Piccolino, 1998)). Galvani's discovery led to the development of Galvinism, or as it known today — electrophysiology, and in the intervening 200 years our understanding of bioelectricity has led to the development of ever more sophisticated ways of detecting and measuring it. As a result we have an advanced ability to measure electrical activity in the brain, but still a comparatively poor ability to detect other processes such as gene expression, calcium dynamics, or neurotransmitter levels. One neuroscientist has gone as far as to suggest that if we had developed calcium imaging before micro-electrode recording then our theories in neuroscience would look very different (Katz, 1999).

But there is plenty of evidence that neural action potentials do correlate with behaviour — some of the most striking comes from experiments done by Wilder Penfield and Herbert Jasper on epileptic patients during surgery (Penfield and Jasper, 1954). By keeping patients under only local anaesthetic they were able to electrically stimulate parts of the cortex to determine which regions were involved in seizure generation. From their patient's reports they were able to sketch out maps of the cortex indicating which regions were involved in which behaviours, and in the process they demonstrated that *electrical stimulation was capable of generating first-person perceptual experience*. This evidence seemed to provide very strong support for the notion that action potentials were the cause of sensory experience, and as a result this work influenced much of what was to follow.

Later work by Mountcastle, and then Hubel and Wiesel appeared to add further support for the belief that sensory perception had its roots in the spikes of neurons. Mountcastle demonstrated by careful recordings that neurons in the cat somatosensory cortex appeared to be organised into microcolumns according to their response properties (Mountcastle, 1957). The fact that neurons were organised in a systematic way seemed to indicate that they were laid out in this way for a purpose, as if the neurons act as components of our perceptual machinery. Nowadays with new recording techniques we know that the same is true of other cells, such as cortical astrocytes (Schummers et al., 2008), but it is easy to see how at the time the evidence seemed to be pointing towards the neuron being the functional unit of the nervous system.

This work was followed upon by Hubel and Wiesel, who went on to receive the Nobel prize for their work on ocular dominance columns in the visual cortex. Hubel and Wiesel discovered that neurons in the primary visual cortex of a cat appeared to be selective for patterns of light and dark in particular orientations. Some neurons, which they termed complex cells, selectively fired action potentials when the stimuli were moving in a particular direction (Hubel and Wiesel, 1962). After their discovery it was proposed that our visual experience is constructed of the collective response of many of these simple and complex cells across the entire visual field. Together these discoveries seemed to pave the way for applying the computational metaphor in neuroscience, using the spiking of individual neurons as the functional unit of computation.

But it is not just experimental observations that have led to the view that spike based models are the most appropriate basis for explaining behaviour — the idea that neuron action potentials are responsible for our perception and action has long been popular amongst theoreticians and empirically minded psychologists. The most famous example of this viewpoint comes from Donald Hebb, who in 1949 proposed a view of the nervous system centred around the function of neurons (Hebb, 1949). He suggested that neurons could act together as cell-assemblies to represent stimuli, and that the relationship between stimulus and response was dictated by the connections between the neurons in these cell-assemblies. He also famously suggested that the relationships between stimulus and response could be learned and remembered according to persistent changes at the synapses between neurons.

This work was of course highly influential, and much of what he proposed in his overarching theory of neuroscience is still assumed to be broadly correct today by experimental and theoretical neuroscientists. The beauty of Hebb's grand theory was that it offered neuroscientists a unifying framework within which all of the experimental

data known at the time could be interpreted and understood.

Since the time Hebb made his prediction about the synaptic basis of memory there has been an ongoing effort to test his ideas experimentally. One particular piece of work which has lent strong support to Hebb's ideas about synaptic plasticity was that done by Eric Kandel on the gill withdrawal reflex in Aplysia. By choosing the simplest possible model organism that demonstrated classical conditioning, Kandel was able isolate the physiological changes involved in the learning process, and demonstrate that learning occurred by synaptic changes (Kandel and Tauc, 1965).

With these pieces of the jigsaw complete, we can now see why spike-based models of behaviour are so widely accepted in theoretical neuroscience. The results of these seminal experiments allow computational neuroscientists to argue that

1. Neural computation implemented by spiking neurons correlates with or is *the cause of* perceptual experience (Penfield and Jasper, 1954), (Hubel and Wiesel, 1962).

2. we learn stimulus-response or *input-output* relationships by means of persistent changes at synapses (Kandel and Tauc, 1965).

Together these two observations form the basis for a description of *perception*, *motor control*, and *learning*, which to some is a complete view of human behaviour.

This view, which can be traced back to Hebb's overarching theoretical model, has support from the results of Penfield, Mountcastle, Hubel and Wiesel, and Kandel. Today much of what we take for granted about neuroscience comes from this work, and it isn't always recognised that much of *this particular model of behaviour is a result of the historical development of the field rather than a reflection of the current state of knowledge*. Although this model has proved incredibly useful in the past 60 years, *there are reasons to believe that this is not the only model which can be used in neuroscience to explain the relationship between the brain and behaviour*. Since this work was published our knowledge about other non-neural processes has increased, as has our ability to record and relate non-neural processes to behaviour.

The development of spike based models of behaviour has been particularly successful in short-timescale models of early-stage perception and action (for example visual perception (Marr, 1983), or place perception (Burgess et al., 2007)), but we have not yet been able to relate spike-based models to high level behaviour and motivational states that are observed over longer periods. Even so many theoreticians still use neurocomputational models to explain high-level psychological phenomena such

as schizophrenia (Olney et al., 1999), motor learning (Marr, 1969), (Albus, 1971), (Ito, 1984), recognition memory (O'Reilly et al., 1998), language (Pulvermüller, 1999), (Smolensky and Legendre, 2006), and attention (Ardid et al., 2007).

60 years on from the publication of Hebb's framework for neuroscience, does it still make sense to use his model to explain the relationship between brain and behaviour? Is there evidence that we have found in the meantime that might suggest an alternative model which could help us make more progress? Should spikes and neural computation remain our biological basis for talking about psychological concepts? Can we use this framework of neurons and spikes to explain the relationship between brain and behaviour?

1.2.5 Summary

In this section I have argued that our choice of models and metaphors in neuroscience has been shaped by the historical development of the field. I have suggested that recent discoveries in neuroscience, and our changing expectations of what neuroscience should do, has not been reflected in the development of new theoretical models. In the next section I will discuss specific evidence which indicates where and how we might want to change our models.

1.3 Beyond neural computation

In the last 20 years theoretical and computational neuroscience has proved quite successful in providing accounts of psychological processes. Computational neuroscience has been most successful in *explaining* low-level perceptual and motor phenomena, but as yet this has not been followed up with *quantitative predictions* of higher-level psychological phenomena. Why is this?

During the early stages of developing my thesis I was interested in the question of how processes in the brain relate to executive behaviour, such as decision making and learning from reward. The more deeply I researched these topics, the more facts I learned about neurophysiology that didn't make sense if we took the view that spike-based neural computation was the basis of executive behaviour. In this section I will discuss a few of these pieces of evidence that indicate the need for models that go beyond current models of behaviour based upon neural computation. I will first discuss evidence that suggests a need for models that go beyond those based upon neurons and

action potentials (Section 1.3.1). I will also make a separate argument as to why we may want to consider the use of metaphors other than computation (Section 1.3.2).

1.3.1 Beyond spike-based models

1.3.1.1 The diversity of neurotransmitters

To start with a basic observation that has been made before (Edelman, 1993), (Leng and Ludwig, 2006) - if it is the patterns of spikes in the brain which are significant, then why is there such a diversity of structures and neurotransmitters in the human brain? When we quantify our models in terms of spike counts we effectively assume that the neurotransmitter involved is unimportant. If glutamate alone can potentially cause any pattern of spikes which code for perceptual stimuli, then why do we need such a zoo of signalling molecules in our brains?

Perhaps the answer to this question is in the ways these signalling molecules differ in their properties and timescales. By introducing a myriad of levels of complexity in the diffusion, binding dynamics, and intracellular cascades that happen simultaneously at multiple timescales, the brain has potentially a much richer way of representing information than if it was only reliant upon point-to-point synaptic transmission.

It is occasionally claimed that diversity of neurotransmitters is a hangover from our evolutionary past, and not actually functionally relevant — this argument will be discussed in section 1.3.1.6.

1.3.1.2 Correlations between neuromodulators and behaviour

Following on from this, it has been known for many years that non-glutamatergic neurotransmitters such as dopamine, noradrenaline, and serotonin are particularly strong correlates of our behavioural state (Carlsson et al., 1957), and have significant effects upon high level behaviour (Schildkraut, 1965). Theories like the monoamine theory of depression have been influential in biological psychiatry since the 1960s, but have had surprisingly little impact in theoretical neuroscience.

Significantly, the same neurotransmitters which appear to have a strong effect upon behaviour are also evolutionary conserved, and re-occur or have close analogues in vertebrate and invertebrates. This seems to indicate that *it is the properties of the neurotransmitters that are functionally important, and not only only the patterns of spikes in which they are released.*

1.3.1.3 Behaviour without spikes

Recent research into simple organisms like Caenorhabditis elegans seem to indicate that neuronal action potentials may not even be necessary for behaviour to occur at all. C elegans is able to demonstrate recognisable behaviour despite the fact that its neurons do not fire action potentials (Lockery et al., 2009).

1.3.1.4 The importance of non-classical neurotransmission

At present much of the work in computational neuroscience is based upon classical neurotransmission, whereby an all-or-nothing action potential causes the release of neurotransmitter vesicles into the synaptic cleft. There the neurotransmitter binds to the post synaptic terminal and causes a depolarisation or hyperpolarisation of the post synaptic neuron. Since this simple model of neurotransmission was developed, neuroscientists have made many observations that appear to contradict it:

1. action potentials are not all or nothing (Alle and Geiger, 2006)
2. where synapses occur, the interactions are often tri-partite rather than bi-partite (Araque et al., 1999)
3. the terminals of many neurotransmitters do not form synapses, and instead rely upon volume transmission (Vizi et al., 2004).

The evidence that has accumulated in recent years for volume transmission does seem to form a challenge to the view that we can understand the brain in terms of all-or-nothing action potentials at synapses. If a neurotransmitter is released into the extracellular space rather than a synaptic cleft then it indicates that the diffusion, reuptake, and binding dynamics are likely to have effects upon postsynaptic neuron responses which could not be captured in spike counts alone. Herkenham (1987), in his survey of neurotransmitter release sites found that co-localisation of release sites and receptors was the exception rather than the rule. This is particularly significant in the case of noradrenaline, acetylcholine, serotonin, and dopamine, where 70-90% of varicosities do not make synaptic contacts despite having the necessary machinery to do so (Vizi et al., 2004). In addition to the complications added by the monoamines, recent years has seen the discovery of nonconventional transmitters such as nitric oxide, carbon monoxide, and hydrogen peroxide. These gaseous molecules can cross biological membranes and diffuse large distances from the site of production making it very

difficult to relate their effects to classical forms of communication within the nervous system (Vizi et al., 2004).

1.3.1.5 The role of non-neural cells

In section 1.2.4 I discussed the historical reasons as to why there is a focus in neuroscience upon neurons as a functional unit of the nervous system. Since those seminal experiments were done, we have gradually accumulated evidence which highlights the importance of non-neural cells in brain function. Despite the strong focus in *neuro*science upon neurons, these cells make up only 50% of the cells in the brain (Azevedo et al., 2009). Today there is an increasing awareness of the active roles that glial cells play in learning and behaviour. However, this knowledge has not yet had an effect upon our theories or computational models.

Recent developments in calcium imaging have shown that astrocytes can be more tightly selective for perceptual stimuli than neurons (Schummers et al., 2008). And the astrocytes that form a part of the tri-partite synapse can have a modulatory effect upon neuron-neuron transmission. Glial cells play an important role in responding to inflammation, and the cytokines that they synthesise and release (Hanisch, 2002) can have a significant effect upon neural behaviour and synaptic plasticity (Schneider et al., 1998). It has even been argued that due to their important role in neural homeostasis, glial cells are the fulcrum of brain pathology (Giaume et al., 2007), as glial cells, when failing to function, determine the degree of neuronal death. If much of the funding for basic neuroscience research is eventually intended to solve public health problems, then perhaps our models should be more focussed towards understanding the role of glial cells in generating behaviour.

1.3.1.6 Evolutionary perspectives on neuroscience

As discussed in section 1.2.4, one of the major reason we believe that spike-based models are a good means for explaining behaviour is that neural activity seems to correlate well with perceptual experience (Penfield and Jasper, 1954), (Mountcastle, 1957), (Hubel and Wiesel, 1962). But it is important to realise that many of the recordings that have confirmed this view were performed in the cortex. Being on the outer surface of the brain the cortex has historically been easier to record from, and this may have affected the type of data neuroscientists have been able to gather.

Also, the cortex is one part of the brain which is thought to be relatively enlarged

in our species (see (Azevedo et al., 2009) for evidence to the contrary), and so has been assumed to be the structure that distinguishes apparently intelligent homo sapiens from other apes and vertebrates.

If these two claims are true, then it suggests that *both our data and our motives contain biases*, and we should be aware of this when constructing models.

From an evolutionary point of view, constructing a model of behaviour based upon activity observed in the cortex is an unusual way to build a model of high level behaviour. Unlike human designed artefacts, organisms that have developed by evolution are constrained in their function by the biological machinery that was present while they evolved. This would imply that *models of high level behaviour ought to be developed within the constraints of our understanding of the most primal, basic parts of our nervous system*, rather than beginning with the neuroanatomy which appears to be most different in our species. If we examine the brain in its evolutionary context we can see that the evolutionary conserved behaviours that we share with other species are primarily regulated by systems in the brainstem and hypothalamus. If high level behaviour is constrained by these systems rather than the other way round, then this indicates that we should first aim for an understanding of how these basic circuits work, before moving on to relate their activity to other, more recently developed regions of the brain.

1.3.2 The computational metaphor

To propose a metaphor is to suggest that there is a one-one mapping between a subject and the object of the metaphor. For the metaphor to be apt, the characteristics and properties of subject and the target must be similar.

Computation as is most often described is a rational, rule-based process which takes one from state A to state B. This would make computation a suitable metaphor for explaining the rational processes, but not such a good metaphor for describing the non-rational or unconscious processes which account for a significant proportion of our behaviour (the exact proportion is an ongoing debate (Freud, 2005), (Skinner, 1976)). If computation is necessarily rule based, then it makes the computational metaphor a weak metaphor for talking about certain regimes of behaviour, such as emotion, where the rational account only comes after an intuitive response. It is notable that computational neuroscience is lacking a good model of these kind of phenomena despite their important role in human psychology. Could it be that the metaphors that we choose

can both *help* and *limit* us in developing models of behaviour?

A computer (and by extension, computation) is a man-made artefact, and in a sense was designed to mimic a particular view of human psychology. As a result, the performance of a computer represents a particular regime of rational human behaviour, and it is not appropriate to apply it as a metaphor in other regimes of behaviour. *Computation represents a school of psychology*, and as such should not be used as a metaphor for psychology itself.

To date the computational metaphor has been most useful in neuroscience when describing psychological processes that can be achieved by computation - for example memory storage, rule-governed behaviour, and feature detection. But there are other important aspects of psychology that do not figure in our normal conception of a computation or a computer - Why do we want or like things? (our drives), and Who/Why am I? (our sense of self). *The models that I will develop and explore in this thesis are models that I believe are better suited to answering these important questions.*

1.4 A separation of timescales

The previous section highlighted empirical and conceptual arguments as to why we might want to explore alternative models and metaphors in neuroscience, but there are also a priori reasons why we might want to reconsider the idea of relating behaviour to neuronal action potentials.

It may seem strange to suggest that there are a priori reasons why we ought to avoid using neural computation as a model for explaining behaviour. How can we use a priori arguments in empirical science?

The argument that I want to make use of is referred to as the "separation of timescales" principle, and is used routinely in the physical sciences as a technique for prioritising the components of a system that are most likely to cause its macroscopic behaviour.

Physicists apply the separation of timescales principle when they have a system with multiple variables changing on different timescales. For example, when trying to explain the dynamics of a simple harmonic oscillator with a frequency of 1Hz, a noisy driving force at 1Hz is much more likely to be causal than a driving force at 100Hz. In this way physicists can disregard phenomena like the 100Hz driving force, and do so in a systematic way, because we know from our intuitive metaphysics that variables changing at microscopic spatial and temporal scales are unlikely to be the proximal cause of macroscopic phenomena. A toy example of how the separation of timescales

timescale	phenomena
nanoseconds	quantal release, ion channel events
microseconds	synaptic proteins
milliseconds	reflex behaviour, action potentials, local field potentials
hundreds of milliseconds	neuromodulation, gene transcription, EEG, glial metabolism
seconds	behaviour, gene transcription, haemodynamics, neurohormones, neuromodulation
minutes	behaviour, gene transcription, neurohormones, stress responses
longer	behaviour, gene transcription, hormones, stress responses

Table 1.1: Timescales of behaviour in neuroscience

argument can be used in neuroscience is shown in Figures 1.1 and 1.2.

The reason why we can ignore these variables is not drawn from the empirical data, but come from our intuitive metaphysical understanding of how scientific explanations ought to look. We know that the principal component of the variation in our data is always due to variables of a slightly smaller spatial or temporal scale than the phenomena we are studying.

This principle is often applied in physics when deconstructing a *natural* system, but the question I want to ask is "what would happen if we applied the principle to neuroscience itself?". What would it tell us about how we ought to construct a explanation of the relationship between the brain and behaviour? Does the model that we currently use in neuroscience — that of explaining behaviour in terms of the combined effect of many action potentials — does this model make sense in terms of the separation of timescales principle? Table 1.1 shows some examples of phenomena in neuroscience, and the timescales at which they vary.

If we assume that the main behaviour of interest in neuroscience is that which correlates with our subjective experience — such as perceptual experience, decision making, etc. — then the small selection of phenomena included here indicates that according to the separation of timescales principle, neural action potentials are a good candidate to be the proximal cause of fast behaviours, and early-stage perception, but not for psychological processes.

In fact there are many other variables we could measure that vary at a timescale

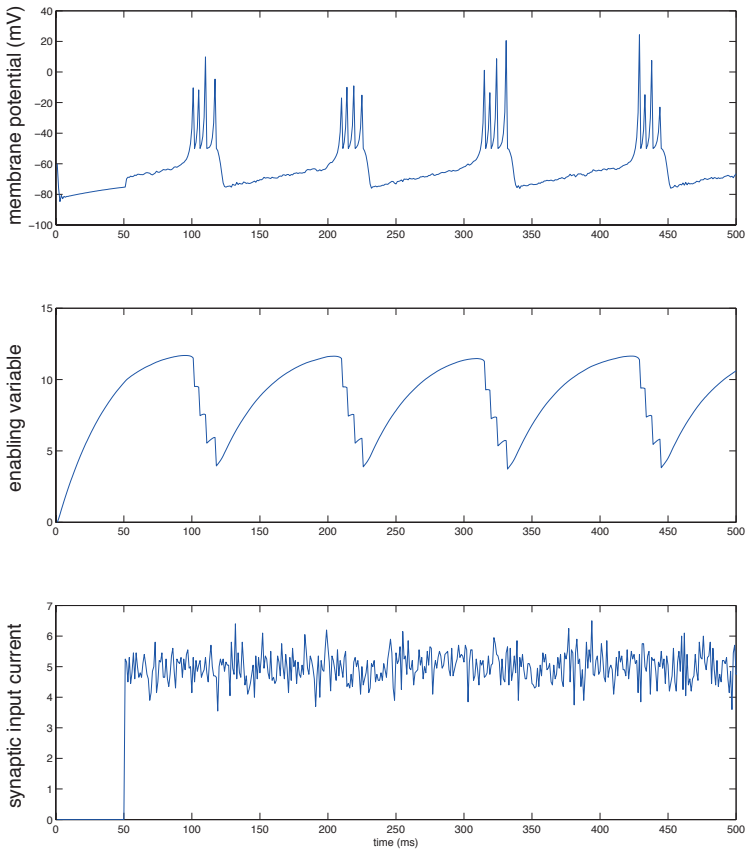

Figure 1.1: A toy model designed to illustrate how the separation of timescales principle can be used in practice. The conclusions drawn from this example are not meant to be taken literally: Suppose we have a neuron which displays bursting behaviour, much like the activity shown in the top graph. The activity of the neuron is governed by two mechanisms - one a quickly varying synaptic input (bottom graph), and another enabling variable, a slower process internal to the neuron (middle graph, think of this as calcium build-up). Now suppose we want to answer the question — "What causes the bursting?" Do we have any a priori reason to believe it is the calcium build-up rather than the synaptic input which causes the spiking? Consult Figure 1.2 for an answer.

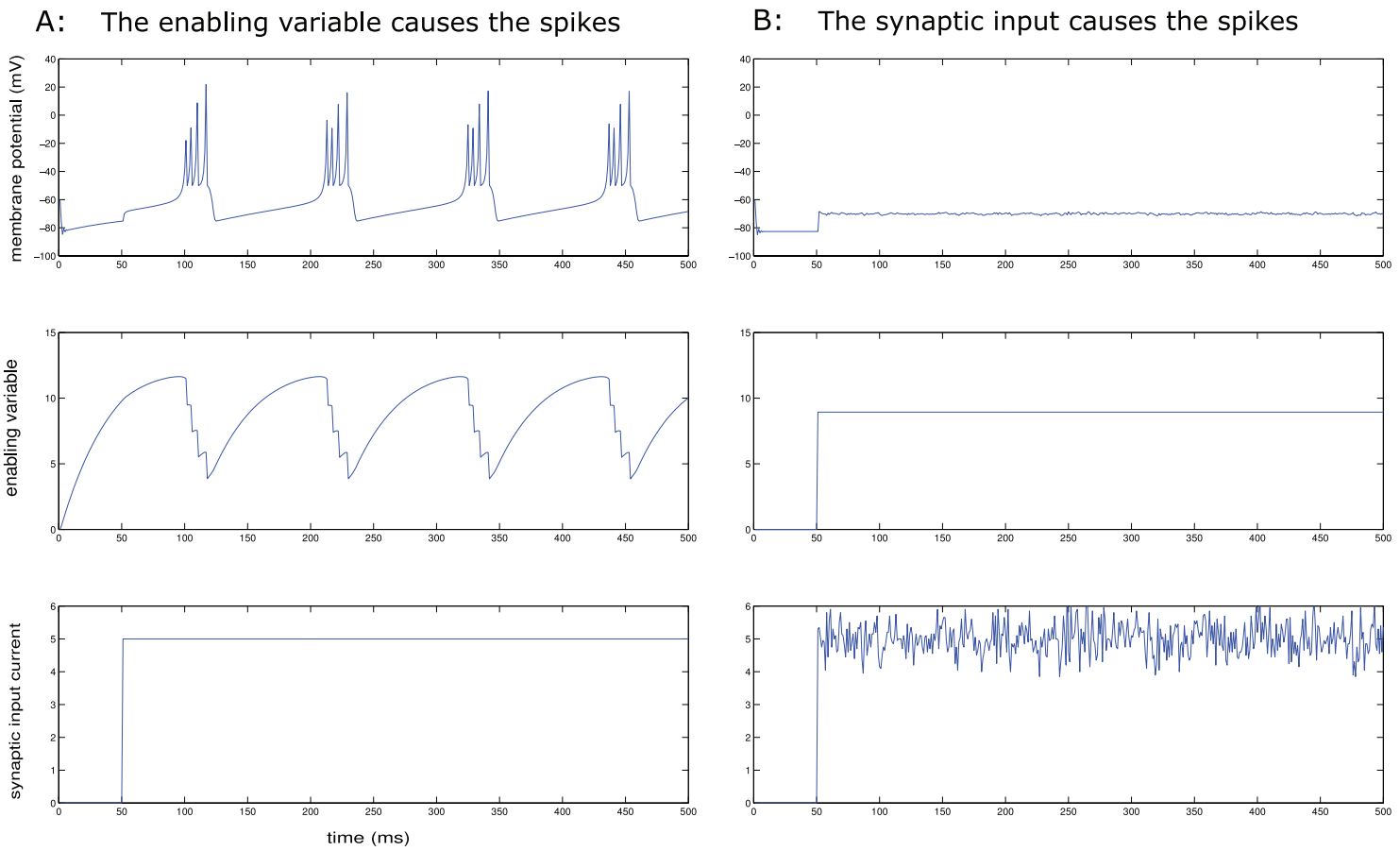

Figure 1.2: Following on from Figure 1.1: Because this is a computational model we can remove the variance from either mechanism (by taking the mean of the time-series) and look at the effect this has on the behaviour of the neuron. On the left hand side we remove the variance in the synaptic input, and on the right hand side we remove the variance in the calcium signal. As you can see retaining the signal provided by the calcium maintains the bursting behaviour of the neuron, whereas retaining the synaptic input does not. *In this sense we can say that it is the calcium build-up which causes the bursting behaviour, and not the synaptic input*. If we had used the separation of timescales principle beforehand we could have predicted this, because while the synaptic input does have some possibly meaningful variance, it is not at the right timescale to be the cause of the bursting. A more technical description could say that the calcium signal is more likely to be the cause of the bursting because the cross-correlation between the two signals is stronger than the cross-correlation between the membrane potential and the synaptic input. This simple toy example should serve to demonstrate a process which commonly goes on in neuroscience. Much like in this example, the theoretical model put forward by Hebb encourages researchers to assume that the synaptic *input* must be the cause of the *output* behaviour, irrespective of the timescales of the output behaviour in question.

closer to the timescales of the behaviour we are interested in (here I am assuming that as neuroscientists we are most interested in psychological processes and their related behaviour).

This argument would suggest that we would be better off constructing models of behaviour from patterns in gene transcription, EEG signals, haemodynamics, or glial dynamics, rather than spikes. The fact that we don't reflects two things:

1. When viewed at the timescale of seconds, phenomena varying on the timescale of milliseconds tend to appear as binary, all-or-nothing phenomena. Discrete, binary variables are easier to quantify than those that change continously.

2. Because gene transcription, EEG signals, and haemodynamics tend to vary more slowly than neural spikes they are difficult to quantify. As a result they are in practice more difficult to measure, and therefore do not make their way into computational models.

Although phenomena like gene transcription and haemodynamics are more awkward to quantify, they are not intrinsically unsuitable components for a model. These kind of phenomena are studied more frequently in subdisciplines such as cognitive neuroscience, but they are often thought of as proxy variables for neural activity, rather than causal in their own right.

There is one particular phenomenon in Table 1.1 which has recently begun to be used as a basis for models in neuroscience - neuromodulation. Developments in the last 40 years has made it possible to measure the concentrations of neurochemicals in the brain in vivo (Robinson et al., 2008), and the data that has come from this is beginning to have its effect on neuroscience, as theorists become more aware of the behaviourally linked fluctuations of neuromodulators that occur in awake, behaving animals.

When beginning this thesis my particular interest was in executive behaviour, decision making, and goal directed behaviour. After spending months of reading the literature it became clear to me that explaining the biological basis of this behaviour in terms of the spikes and computations of individual neurons didn't make much sense from the perspective of the argument outlined above. There has been an awareness for a long time that drugs which affect neuromodulation have a strong effect on executive behaviour, and recent developments in experimental techniques are providing increasing amounts of evidence that does not fit well with the view that executive behaviour comes about as a result of neural computation.

One example of these experiments is the discovery that dopamine neurons fire in ways which correlate with theories about decision making. Recording of dopamine neurons by Wolfram Schultz have shown that these neurons appear to fire en masse when an animal encounters an unexpected reward, thereby simultaneously releasing dopamine across disparate areas of the brain (Schultz, 1998). This is particularly interesting as the unexpected reward signal closely mimics the reward prediction error variable that had previously been proposed by abstract models of learning (Schultz et al., 1997). Observations like these seem to point a way towards alternative models of behaviour whereby the active components are neuromodulators rather than neurons.

This accumulation of evidence, and the difficulty in reconciling it with the way in which most neuroscientific explanations are currently made led me to become interested in the bigger question — is a model of behaviour based upon the actions of neuromodulators a more appropriate framework than one based upon the neural computations of spikes?

1.5 Neuromodulation as a means for explaining behaviour

1.5.1 A definition

Neuromodulation is the process by which a neurotransmitter released into extracellular space affects the ongoing spiking activity of one or many neurons. This is in contrast to classical neurotransmission whereby one presynaptic neuron directly influences a single postsynaptic partner. Dopamine, noradrenaline, and serotonin are common examples of neurotransmitter than can act as neuromodulators, but there is a huge number of lesser known signalling molecules and peptides that can modulate neural activity.

Sherman and Guillery (1998) introduced a distinction between "drivers" and "modulators" — inputs to a cell that can either affect neuron activity on a short timescale (the spiking activity in their example), or alter activity over longer periods (through changes to the neuron excitability in their example). A more technical definition would state that a "driver" tends to have a narrow peak in the cross-correlogram between the input and the output, whilst a "modulator" may have a lower but wider cross-correlation peak. An example of this is shown in Figure 1.3

In this sense any neurotransmitter or signalling molecule can be understood as a modulator relative to any ongoing process that it affects — for example glutamate may

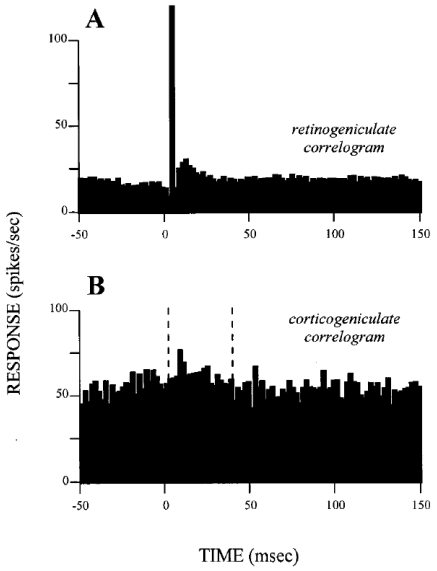

Figure 1.3: Adapted from (Sherman and Guillery, 1998). Graph A shows the cross-correlogram typical of a driver input. In this case it is the activity of the retinal neurons driving activity in the geniculate. Graph B shows the effect of cortical neurons upon the geniculate. This cross-correlogram is characteristic of a modulator.

be considered a modulator of gap junction transmission in astrocytes. Defining neuromodulation with this degree of flexibility can be useful when we want to integrate neuromodulation with other important processes in neuroscience, such as those mediated by the endocrine and immune systems. By defining neuromodulation in such a broad sense we can describe endocrinological effects on the system using the same metaphor — we can describe neurohormones as modulating the neuromodulators. For an examples of this, see the effect of steroid hormones or orexin on the firing of dopamine neurons (Mesce, 2002), (Korotkova et al., 2003).

1.5.2 How do neuromodulators have a consistent effect on behaviour?

If neuromodulators are released into the extracellular space and allowed to diffuse across the brain, you might wonder how such an apparently non-specific signal can have such a reliable effect upon behaviour. This is a question that researchers have been trying to answer for many years, and much of the work done to answer this question has taken place in crustaceans.

Neuroscientists have chosen to focus their studies on neuromodulation in crustaceans because the neural wiring of these organisms is identical across different in-

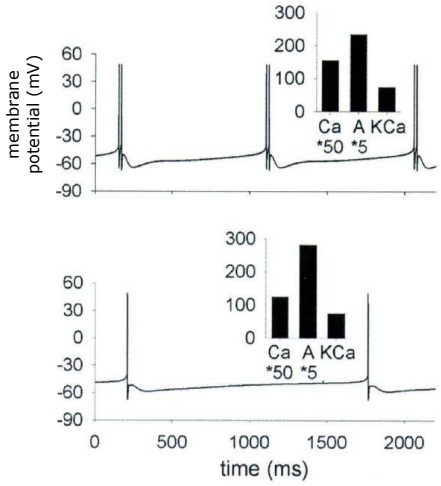

Figure 1.4: Figure adapted from (Goldman et al., 2001). Very small changes to ion-channel parameters in a neuron model can result in dramatic changes to the spiking behaviour. It has been proposed that neuromodulators might cause reliable changes in neural behaviour by acting at such phase boundaries to trigger a bifurcation in the dynamics of the neuron.

dividuals of the same species. This allows whole circuits to be mapped out, and the effect of individual neuromodulators on specific cells to be assessed. In crustaceans neuromodulation of central pattern generating circuits in the stomatogastric ganglion have been studied in particular detail (Marder and Bucher, 2007).

It is thought that the ion channel-specific effect of some neuromodulators is enough to allow them to affect neural circuits in a deterministic way. Small changes to a specific type of ion channel in a circuit of interacting neurons can be enough to cause a switch in the qualitative behaviour of a circuit (Goldman et al., 2001). For an example of this see Figure 1.4

The authors also proposed that these small changes in ion channel parameters might occur at boundaries in the neuron's state space, such that the small change in parameters triggered by the neuromodulator is enough to cause a bifurcation in the dynamics of the circuit, and result in a switch to a qualitatively different pattern of firing. In this way a single neuromodulator released by volume transmission and allowed to diffuse across the brain can have a coherent and reliable effect each time it is released.

1.5.3 The effect of neuromodulation upon behaviour

In the human brain most neuromodulators are synthesised by nuclei situated in the midbrain and hypothalamus. Some of these nuclei such as the ventral tegmental area (dopamine), the raphe nuclues (serotonin), or the locus coeruleus (noradrenaline), to name but a few, contain neurons which project from the midbrain into the striatum and cortex, and it is here they are thought to exert their influence on higher level behavioural processes. From recent experiments we are now aware that there are precise patterns of neuromodulator receptor distribution in the cortex and striatum (Hurd et al., 2001), (Surmeier et al., 1996), and it may be that it is *the degree of innervation of these axons and the receptor distribution that combine to determine the firing patterns of cortical neurons* that we had previously thought was a result of the neurons' intrinsic properties.

A good example of how strong an effect a non-specific neuromodulatory signal can have on behaviour comes from studies on the pair-bonding behaviour of prairie voles. A study by Winslow et al. (1993) found that it is the action of vasopressin which causes the selective aggression and pair-preference behaviour of male prarie voles. Male prairie voles that were injected with a vasopressin antagonist 24 hours before sexual experience failed to demonstrate a partner preference or selective aggression towards other males 48 hours later. Males that were injected with CSF or an oxytocin antagonist showed no change from the normal behaviour. This experiment demonstrates quite clearly that such psychologically fundamental behaviour as mate choice and social bonding can be easier to explain using neuromodulation than it would be using patterns of action potentials.

1.6 Using neuromodulation to explain decision making

As we can see, there is considerable evidence to suggest that neuromodulation might be a reasonable framework for trying to explain the physiological basis of behaviour. However, one of the major advantages of basing models upon neural action potentials is that it is quite clear how they can be quantified and used to build computational models. How can we describe the effects of neuromodulation in a quantitative way? One option might be to use reinforcement learning.

As I mentioned in section 1.4, recent work by Montague et al. (1996) has shown that dopamine neurons fire in ways that correlate with reinforcement learning theory

- an abstract framework developed by engineers and mathematicians to describe the optimal way to make decisions. In particular dopamine neurons correlate very well with the reward prediction error signal proposed by the Temporal Differences (TD) algorithm — an implementation of reinforcement learning that uses the differences in prediction at successive timesteps to drive the learning process. The observation that dopamine neurons correlate with with the reward prediction error is quite an incredible discovery — that neurons in the midbrain precisely mimic the predictions of abstract model of decision making.

But this raises and obvious question — why does this happen? One answer to this question, and an answer that has generated much research interest is that the *dopamine neurons are firing in accord with the temporal differences algorithm, because they are implementing the temporal differences algorithm*[1].

This claim, apart from its boldness, points a way towards how we might develop quantitative models of neuromodulation: *By attempting to find neural correlates of the variables in the TD algorithm we can assess whether or not the brain really does use dopamine to implement reinforcement learning.*

Suppose for a minute that this dopaminergic reinforcement learning hypothesis is true, how do we accommodate the fact that many other neuromodulators also seem to affect decision making (Robbins, 2005)? How do we make sense of the fact that dopamine neurons themselves are modulated (Korotkova et al., 2003) — what do these meta-modulators represent in the Temporal Differences (TD) model?

One author has gone so far as to suggest that the major neuromodulators which innervate the cortex act as the different parameters of the TD algorithm (Doya, 2002). In Doya's model he suggests that dopamine signals reward prediction error, serotonin controls the timescale of reward prediction, noradrenaline the randomness in action selection, and acetylcholine the speed of memory update.

The initial paper linking dopamine and the TD algorithm has had a significant impact in the field, and since that time many others have built upon this model by analysing the dynamics of a particular neuromodulator, and proposing a role for it

[1]The reader may have noticed that I described this implementation of reinforcement learning as the temporal differences *algorithm* — the use of "algorithm" suggesting that it is a computational process. It is true that that reinforcement learning is compatible with the computational metaphor and therefore is not necessarily a replacement for it. In fact many of its proponents do describe it in terms of computation. However, reinforcement learning need not be described in terms of computation, and it is for this reason that I consider it a potential alternative model. Reinforcement learning can potentially be interpreted using the metaphor of feedback described by cybernetics (Rosenblueth et al., 1943). Or it could also be described in the language of autopoesis, as a system that creates its own environment in which it is optimal (Varela et al., 1974).

as a signal in a reinforcement learning circuit. These suggestions have ranged from acetylcholine signalling expected uncertainty (Yu and Dayan, 2005), noradrenaline signalling unexpected uncertainty (Yu and Dayan, 2005), serotonin mediating behavioural inhibition (Dayan and Huys, 2008), noradrenaline mediating the exploitation:exploration trade-off (Aston-Jones and Cohen, 2005), and dopamine signalling unpredicted sensory events for which the organism is responsible (Redgrave et al., 2008).

This process is intriguing as theorists step in to fill in the gaps of a newly developing model. There is a clear similarity between this nascent model of behaviour and drive theory in psychology (Hull, 1943) - *rather than explaining behaviour in terms of computation, neuroscientists are beginning to explain behaviour in terms of desire*.

Together these developments indicate that a framework is shaping up to explain behaviour (albeit decision making behaviour in this case), in terms of the action of neuromodulators rather than patterns of action potentials. However, this is a relatively new proposal, and *it is unclear whether or not the model is supported by empirical data*.

At present we know that the firing pattern of dopamine neurons closely matches the reward prediction error signal of the temporal differences algorithm when tested in laboratory conditions, but how and why is dopamine released in this way? We believe that this reinforcement learning model may map onto the actions of neuromodulators, but at present we have no mechanistic model of *how* they might achieve this.

My aim in this thesis will be to explore the developing mechanistic models that attempt to put flesh on the bones of this theory. Does what we know about the physiology match up with how reinforcement learning models work?

In this thesis I will try to bridge the divide between belief in the theoretical model and the evidence provided by the biological data. I will do this by constructing models that allow us to compare how we think neuromodulation works, with what we find when we do experiments. If I formalise these conceptual frameworks as computational models, will the models produce data that is consistent or inconsistent with the hypothesis? Can I use computational models to support or refute this hypothesis?

1.7 The questions in this thesis

To summarise, in this thesis I will attempt to address 3 parallel questions:

1. Does dopamine release form part of a reinforcement learning circuit in the brain?

2. Does the dopaminergic reinforcement learning model indicate that we can explain the relationship and behaviour better by looking at the brain from the perspective of drives and neuromodulation, rather than computation at the level of single neurons?

3. Can we settle these empirical questions using computational models?

In each chapter I will formulate ways in which these high level questions can be addressed using computational models (See sections 2.6, 4.2.3).

The reasons for asking the first two questions have already been covered in this chapter, but I have added the last question because I believe that justifying my methodology is an important part of the thesis. I am sceptical about what computational modelling can achieve, and this view is shared by many experimental neuroscientists. Do these models tell us anything new, or are they just formalised thought experiments? My hope is to justify (at least for myself), that these models can play a part in neuroscience.

During the course of this thesis I will also describe the theoretical and practical problems I encounter with the models I construct. Hopefully by the end of the thesis I will have done enough to justify that my means of addressing this hypothesis, and therefore that my conclusions regarding the first two questions, will be valid.

1.8 Thesis plan

Chapter 2 contains a literature review of the empirical evidence for the involvement of dopamine in a reinforcement learning circuit. There is also some background on reinforcement learning to provide some of the terminology that is used in later chapters.

Chapters 3, 4, and 5 make up the computational modelling work done during the course of the thesis.

Chapter 3 covers work done to reimplement and analyse a model of dopamine-modulated classical conditioning. In analysing the model it was found that some of its behaviour was incompatible with it being a true model of classical conditioning. From the analysis it is concluded that problems encountered in the model were a result of it being constructed from theoretical models rather than empirical.

As a consequence, the aim of Chapter 4 is to develop an empirically derived model. This model aims to investigate some of the assumptions of the model in Chapter 3 by constructing a data-driven model of dopamine modulation of synaptic plasticity.

However, in the process it is found that even with this empirically based model it is difficult to incorporate all the relevant observations. The model produces results which suggest that the synaptic effects of dopamine are more significant than the effect of dopamine on excitability.

Following Chapters 3 and 4 it is decided to make a third model which aims to be a clear conceptual model, rather than one which is verifiable by the empirical data. This model is described in Chapter 5 and is found to be able to successfully implement a reinforcement circuit, and backpropagate reward prediction error.

Chapter 6 contains a discussion of the findings from the computational modelling work. The findings from the models were inconclusive, and so I examine the remaining motives for accepting the dopaminergic reinforcement learning hypothesis. To put the results in a greater context I include some data which does fit well with the hypothesis.

In Chapter 7 I review the transmitter systems that modulate dopamine release and use them as a basis for exploring the context of dopamine modulation and reward-driven behaviour. I look at the function of dopamine from the context of stress and immune responses, and suggest alternative ways of interpreting the function of the dopamine system.

Chapter 8 contains the conclusions of the thesis.

Chapter 2

Dopamine and Reinforcement Learning: A literature review

2.1 Novel contributions

This chapter reviews the literature on computational models of dopamine modulated reinforcement learning. It also integrates the existing empirical data on the biophysical effects of dopamine modulation from the anatomy and firing dynamics of dopamine neurons, through to the diffusion and binding of dopamine, and the eventual intracellular and electrophysiological effects of dopamine.

2.2 Introduction

This chapter begins with a brief historical account of how the neurotransmitter dopamine came to be linked with reinforcement learning. After this historical account I will review the literature that could be used to construct a computational model of *how* dopamine might implement reinforcement learning.

2.3 The dopaminergic reinforcement learning hypothesis

The neurotransmitter dopamine was first identified as a transmitter in its own right by Arvid Carlsson and co-workers in 1957 (for a review see (Björklund and Dunnett, 2007b)). It was identified as a new neurotransmitter when Carlsson and his group

found that dopamine was not merely a chemical precursor in the synthesis of noradrenaline, but was being produced and released by other neurons independently of noradrenaline. They were also able to show that the pharmacological agent reserpine worked in part by blocking dopamine receptors — they did this by demonstrating that reserpine-induced cataplexy (a loss of muscle tone) could be recovered from by restoring dopamine and not noradrenaline levels (Björklund and Dunnett, 2007b).

The similarity between reserpine-induced cataplexy and Parkinsonian akinesia, and the fact that both could be reversed by an injection of dopamine precursor L-DOPA, led some researchers to speculate that the new neurotransmitter dopamine was involved in the initiation of movement and motor control. Later studies supported this view as it became clear that dopaminergic cell death was a major factor in the development of symptoms in Parkinson's disease.

In the 1970s Wolfram Schultz set out to investigate why dopaminergic cell death resulted in Parkinson's disease. Schultz recorded from dopamine neurons in the brains of monkeys, hoping to find cells involved in the production of movement, but he was unable to find what he was looking for (Lehrer, 2008).

Initially he considered these experiments a failure, and it was only after years of research that Schultz began to notice strange patterns in the way these dopamine neurons fired. Rather than firing when the monkey began to initiate a movement, the neurons seemed to fire just before the monkey was given a reward. At first these rewards had just been used to encourage the monkey to move, but as it became clear that this was what the neurons were responding to, he was able to set up more complex experiments to tease apart how and why dopamine neurons were firing in this way.

Schultz found that the way these neurons fired changed over time as the animal learned to complete the task for which it was rewarded. The way in which these neurons fired was quite complex, and although Schultz realised he was unravelling the brains reward circuitry, he did not know how and why the neurons fired in the patterns they did. Schultz published his findings in late 80s and early 90s (Schultz, 1986; Schultz et al., 1993), but at that time it was still not known why neurons in the midbrain appeared to represent such an abstract concept as reward. Some examples of how dopamine neurons respond when presented with cues and rewards are shown in Figure 2.1.

It was 1991 before theorists working on the Temporal Differences (TD) formalism of reinforcement learning came across Schultz's results and were finally able to explain the patterns he had observed. Reinforcement learning theory was developed by math-

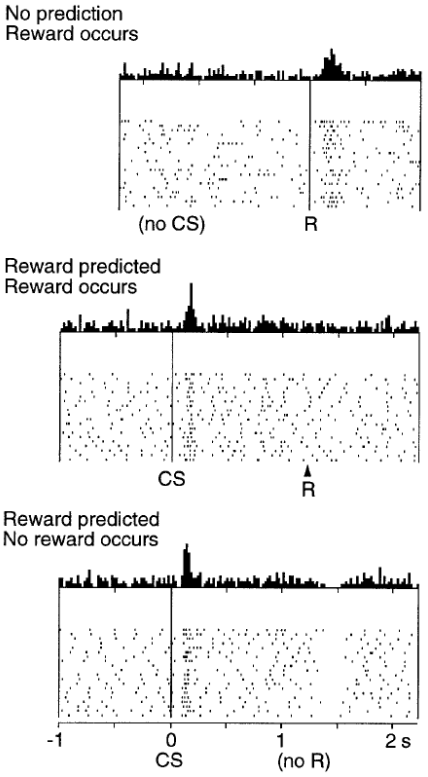

Figure 2.1: Spike raster plots of the effect of a reward (R) on the firing of dopamine neurons across multiple trials. In the top plot, when an unexpected reward is presented, dopamine neurons fire phasically 100ms after the reward delivery. In the middle plot, when a cue (CS) that is known to predict a reward is presented, this triggers phasic firing of the dopamine neurons shortly after the cue presentation. If a cue is presented but no reward arrives the tonic firing of dopamine neurons is inhibited at the time the reward was expected to arrive. Figure taken from (Schultz, 1998)

ematicians, computer scientists, and engineers in response to the question — "What is the optimal way for an agent to learn from interactions with its environment?" The TD algorithm, which grew out of reinforcement learning, proposed that the optimal learning signal would be one which signalled the error in the reward prediction. The algorithm is described in more detail in section 2.4.1.

The theorists who made the connection between the firing of dopamine neurons and the temporal differences algorithm published their model in 1996 (Montague et al., 1996), and since then more data has been collected to support this basic observation (Schultz, 2007, 2002).

But what is the significance of this apparent correlation between an abstract model of learning and the firing patterns of dopamine neurons? How can neurons in the midbrain signal something as abstract as reward prediction error? In the last chapter we looked at some of the many claims that have been made about neuromodulators that play some role in TD (Doya, 2002), (Yu and Dayan, 2005), (Dayan and Huys, 2008), (Aston-Jones and Cohen, 2005), but none of these models explain *how* neuromodulators achieve their purported function.

If dopamine does signal reward prediction error, where does it send this signal to? And how does the recipient receive and act upon that signal? To answer this question we need a mechanism.

The aim of this chapter will be to review the literature that could be used to build a mechanistic model of how dopamine contributes to a reinforcement learning circuit in the brain.

2.4 Reinforcement learning theory

2.4.1 The Temporal Differences algorithm

Before I begin to review the literature on dopamine modulation, I want to briefly outline the key features of the Temporal Differences (TD) algorithm, which dopamine release is supposed to mimic. The temporal differences algorithm is an approach to learning from reinforcement which uses the differences in prediction at successive timesteps to drive the learning process. The algorithm can be described using the following equation:

$$\delta(t) \quad = \quad r(t) + \gamma V(t+1) - V(t) \tag{2.1}$$

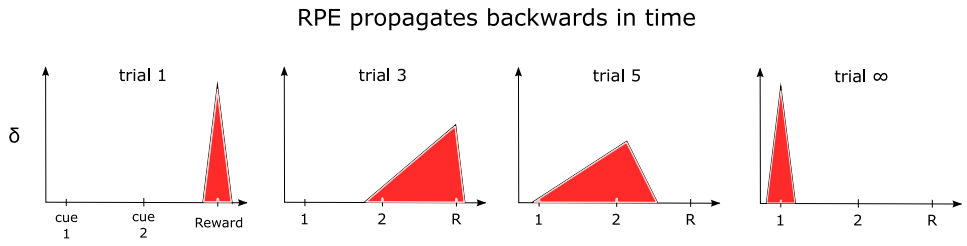

Figure 2.2: A simple idealised schematic of the backpropagation process. In the first trial, the peak in $\delta(t)$ or the RPE (Reward Prediction Error) occurs at the time the reward is presented. With repeated trials the animal's certainty about the reward arrival propagates backward through the delay period (ie. to the left). When the task has been learned the peak in reward prediction error will occur with the presentation of the first cue, which signals without warning that a reward is coming.

Where $\delta(t)$ is the reward prediction error. $r(t)$ is the actual reward at time t, $\gamma V(t+1)$ is the reward predicted at the next timestep, and $V(t)$ is the reward expected at this timestep. In the temporal differences formalism the value of $\delta(t)$ is used to update the internal model, and so it is error in the prediction of reward that drives learning. It is this δ variable that dopamine neurons are thought to mimic. If a reward arrives that is not predicted or expected δ will be high, triggering updates in the model.

This can be seen as a restatement of the Rescorla-Wagner learning rule (Rescorla and Wagner, 1972).

If we implement this model in a predictable environment, then δ will equal zero at each timestep as the prediction is cancelled out by an expectation. But if a reward introduced to the environment, an unpredicted delivery of the reward will result in a transient increase in δ. It is dopamine, or δ that triggers updates in the internal model (ie. learning), and so if a reward is repeatedly presented it could lead to the model predicting the reward based upon any cues that reliably precede it.

As the ability of the model to predict the reward gradually increases, so does an expectation that the reward will arrive. This means that the peak in δ will gradually move backwards in time to the time of the first predictive cue.

This process of the peak $\delta(t)$ moving backward from the time of reward to the time of the first predictive cue is referred to as backpropagation of reward prediction error, and is illustrated in Figure 2.2.

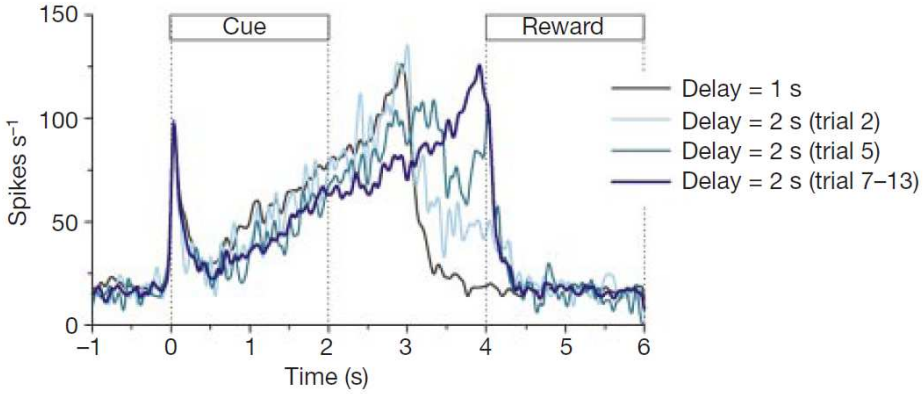

Figure 2.3: A figure taken from (Komura et al., 2001), showing how the ramping activity of neurons in the delay period can quickly adjust itself to reflect the expected time of delivery of the reward. These results indicate that the activity we see in the delay period is shaped by reward delivery, implying that some backpropagation of reward through the delay period must occur.

2.4.2 Backpropagation of reward prediction error

Like the TD variable $\delta(t)$, dopamine neurons also appear to exhibit backpropagation, and like $\delta(t)$ it is thought that dopamine has effects upon learning that cause the update of the brain's reward model. This process of backpropagation is significant because it can potentially explain patterns of neural activity that are learned during the delay period. It has been observed in many experiments that neurons do not only signal information during the presentation of the stimuli, but they also appear to maintain persistent or patterned activity during the delay period (Shuler and Bear, 2006). One example of this is the ramping activity that adjusts itself to peak at the time of the reward delivery (Komura et al., 2001).

Usually this phenomenon is treated separately, but these patterns of activity can be easily explained if we assume that once a task has been learned, the neural activity during the delay between cue and reward is just a consolidation of the activity which was required during training to get the reward. *If particular pattern of (motor) neuron firing repeatedly occurred during the delay period of rewarded trials then it will come to be the behavioural response*. Often this will take the form of persistent cue-related activity during the delay period ie. working memory (Fuster and Alexander, 1971).

Working memory is not often linked with reinforcement learning, but a key claim I would like to explore in this thesis, is that algorithms like TD, may show how working memory may emerge from a reinforcement learning circuit. If true this would

be a powerful result, as it demonstrates how both optimal motor control and flexible, reward related working memory may emerge from a single theory. The obvious parsimony of backpropagation suggests that any model of reinforcement learning ought to implement it.

2.5 Existing biologically inspired models of reinforcement learning

The first work that proposed a link between abstract models in machine learning and the dopamine system was Houk et al. (1994), which suggested that dopamine modulated "striosomes" in the striatum formed part of a circuit that implemented TD-like behaviour. In their model they mapped the actor-critic architecture from machine learning onto the basal ganglia, suggesting that the striatum performed as the actor, and dopamine neurons, the critic. A later model by Suri and Schultz (1998) also implemented this actor-critic architecture in the basal ganglia. A later paper has more explicit predictions of the biological correlates (Suri, 2002), suggesting that reward prediction occurs through corticostriatal transmission, in line with evidence from Hollerman et al. (1998), who found reward expectation activity in the striatum.

Another biologically inspired models of reinforcement learning was proposed by Contreras-Vidal and Schultz (1999) who based their model on Adaptive Resonance Theory (ART), and claim to offer a fuller account than TD-like models. Berns and Sejnowski (1998) offer a systems-level model of how the basal ganglia might implement an action selection circuit, and more recently Izhikevich (2007) claims to solve the distal reward problem by a dopamine modulated spike-timing dependent plasticity model.

2.6 Mechanisms underlying dopamine modulated reinforcement learning

Often when we build models of behaviour in neuroscience we start by examining the dynamics of neurons. This is partly for historical reasons, as proposed in the introduction to this thesis, but also because crudely speaking, it is the firing of motor neurons that determines whether or not an action will occur. If motor control is learned through a process of reinforcement learning, then *a mechanistic model of reinforcement learn-*

ing must at some point explain how dopamine affects (directly or indirectly) the neurons that project to the muscles. There are of course many other processes that play an important role in modulating motor control, but we are not focussing upon these here.

Alongside the need to characterise the effect of dopamine upon motor projecting neurons, I also need to look at *the effect of dopamine upon the systems that feedback to dopamine neurons, and are therefore capable of promoting or inhibiting future dopamine release*. Looking at how dopamine affects neuronal feedback to dopamine neurons themselves is important because this effectively closes the circuit — *if dopamine really does signal reward prediction error, then the effect of dopamine should be to ensure that dopamine neurons signal reward prediction error* - ie. the system must be self-sustaining.

In summary, my mechanistic model of the effects of dopamine will take into account:

1. The effects of dopamine upon motor control neurons
2. The effects of dopamine upon neurons that feed back to dopamine neurons

In order to build this mechanistic model I will review the literature of what is known about each step in the process — from the release of dopamine to its eventual effect upon motor and feedback neurons. When dopamine is released, it diffuses and is subject to reuptake mechanisms, before binding to receptors, influencing intracellular cascades, and then finally affecting the electrophysiology of target neurons.

In this literature review I will focus in particular upon two regions innervated by dopamine — the prefrontal cortex, and the striatum. Both of these areas are innervated by dopamine axons, and contain the neccessary receptors and transporters for dopamine to have a significant effect upon neurons in these regions. I will focus upon the striatum because it contains the richest innervation of dopamine in the brain, is rich in receptors, and most importantly projects back to dopamine neurons themselves. The inhibitory projections that come from medium spiny neurons (MSNs) in the striatum are uniquely placed to influence the firing of dopamine neurons through inhibition and disinhibition. In addition I will also focus on the cortex because, although it has a lesser innervation of dopamine in the striatum, it is known to be involved in motor control, planning, and working memory. In this sense, the cortex may be a key region in generating the behaviour that is to be learned. Neurons in the cortex also send inhibitory and excitatory projections back to the dopaminergic nuclei. Between these

two regions, we potentially have the mechanisms we need to develop a model of how a reinforcement learning circuit may be implemented in the brain.

It is important to note that by focussing upon the effect of dopamine on neurons we may overlook other important non-neural mechanisms - if dopamine has an effect on astrocytes which feedback to the dopamine neurons then this would not appear in a model constructed from the current literature. This may sound a trivial point, but it has recently been demonstrated that the effect of noradrenaline in the cortex is in part derived from its direct effect upon calcium dynamics in astrocytes (Bekar et al., 2008). It is possible that similar effects may also occur with dopamine. In summary, there are enough unknowns about the effects of dopamine release that we should avoid making models that are too dependent upon on only what is known in the present literature.

As a theoretical neuroscientist, by necessity I must make use of data that has already been collected. In the case of dopamine, the bulk of data that exists has been collected to quantify the effects of dopamine release upon neurons.

In the next section I will review the literature and outline what is known about the way in which the proposed reinforcement learning circuit works. The circuit can be broken down into parts

1. The anatomy of dopamine neurons
2. The dynamics of dopamine release
3. Dopamine diffusion and re-uptake dynamics
4. Dopamine receptor binding dynamics
5. Intracellular cascades
6. Electrophysiological effects on neurons

2.7 Literature review

2.7.1 Dopamine anatomy

There are nine major dopaminergic cell types in the mammalian brain, distributed from the mesencephalon to the olfactory bulb. The principle projection neurons of this group are the retrorubral area (A8), the substantia nigra (A9), and the ventral tegmental area (A10), which project to cortical, limbic, and striatal areas. An illustration of the main

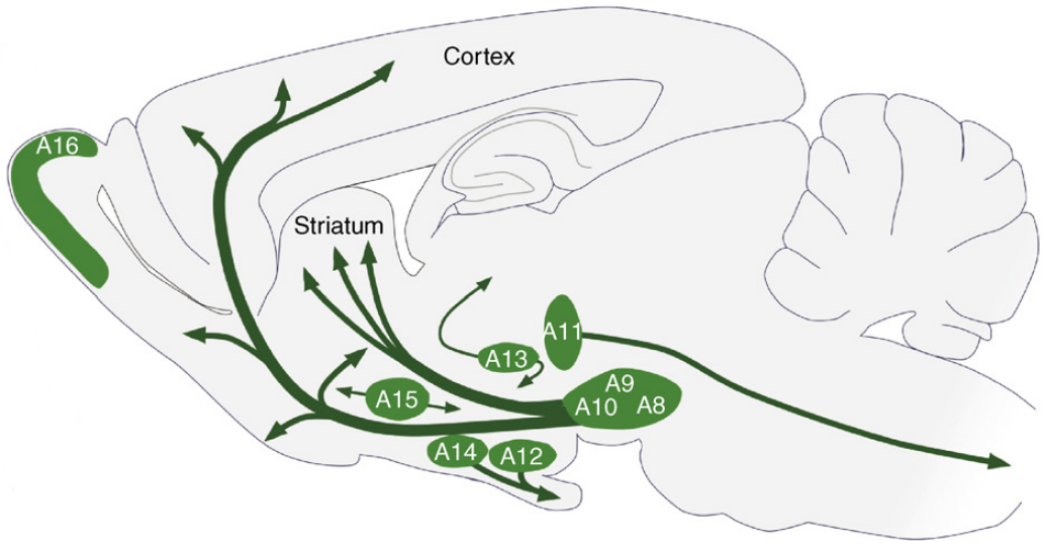

Figure 2.4: An illustration of the location and projections of dopaminergic nuclei in the adult rat brain (Björklund and Dunnett, 2007a).

dopaminergic projections is shown in Figure 2.4. The system is considerably larger in size and complexity as we move from rodents (20-30,000 in mice and 40-50,000 in rats), to primates (160-320,000 in monkeys and 400-600,000 in humans) (Björklund and Dunnett, 2007a).

Although the ventral tegmental area and substantia nigra pars compacta are often described as anatomically distinct and projecting to separate areas, there is a large degree of intermixing, with both the striatum and cortex receiving innervation from the A9 and A10 (Björklund and Dunnett, 2007a). Dopaminergic axons innervate the structures of the basal ganglia, with the striatum receiving by far the densest dopaminergic input. Whilst in the cortex, dopamine axons innervate motor cortical, prefrontal, and anterior cingulate areas, with lesser projections to the primary sensory cortices (Seamans, 2007).

Dopamine neurons in the A9 and A10 regions are often described as showing tonic and phasic modes. During normal waking behaviour in the absence of rewarding or salient stimuli, dopamine neurons exhibit tonic pacemaking activity, with neurons spiking asynchronously at around 5Hz (Arbuthnott and Wickens, 2007). When rewarding, unexpected, or salient stimuli are present, dopamine neurons are capable of phasically spiking en masse, with neurons firing at around 80Hz for 40ms (Arbuthnott and Wickens, 2007). Phasic bursts, which are thought to signal a reward prediction error, can be triggered by glutamatergic afferents from the cortex (Gariano and Groves, 1988), or

by subcortical inputs such as the subthalamic nucleus (White, 1996). Input from the lateral dorsal tegmentum and pendunculo-pontine tegmentum are particularly efficient at triggering phasic bursts, and do so by releasing a combination of acetylcholine and glutamate (Grace et al., 2007; Floresco et al., 2003).

Around 70% of the synaptic input to the substantia nigra is inhibitory, and this input is capable of suppressing phasic bursts. This input comes from many sources, chief among them the striatum and the substantia nigra pars reticulata (Diana and Tepper, 2002), (Celada et al., 1999).

It should be noted that the substantia nigra and ventral tegmental areas receive input and project to a huge number of areas, and so the picture we have presented here is a deliberately simplified one. Diana and Tepper (2002), and White (1996) describe how serotonergic input from the dorsal raphe and noradrenaline from the locus coeruleus can impact dopaminergic cell firing. An exhaustive list of regions with input to the ventral tegmental area can be seen in Figure 2.5.

Alongside the ability of GABAergic input to silence phasic spikes, dopamine-dependent autoinhibition can also act to decrease firing. As dopamine neurons fire they release small amounts of the transmitter from their dendrites, which can bind to D2 autoreceptors, and cause a decrease in the spike rate (Diana and Tepper, 2002).

Although dopamine neurons are so named because of their ability to synthesise and release dopamine, it has recently been suggested that they also co-release several neurotransmitters. Indeed dopamine neurons in the arcuate nucleus (A12) have been shown to colocalise various neuropeptides, such as GHRH, neurotensin, galanin, enkephalin, and dynorphin (Björklund and Dunnett, 2007a). There is also evidence that A9 and A10 projection neurons also co-release CCK, neurotensin, cannabinoids, serotonin, and glutamate in addition to dopamine (Seutin, 2005). The glutamate release from dopamine neurons is significant because it could explain some of the short-latency effects of dopamine neuron firing, which are too fast to occur through dopaminergic pathways (Lapish et al., 2007). In fact it may turn out that some of the most important properties of dopamine neurons needed for reinforcement learning are provided by their release of transmitters other than dopamine.

Questions about neurotransmitter release aside, the key property of dopamine neurons for our purposes is that they appear to fire in ways which mimic the reward prediction error signal in reinforcement learning. The particular combination of afferents shown in Figure 2.5 is somehow enough to cause the neurons to spike in a way which obeys a huge range of well known observations in classical conditioning (Schultz,

Retrogradely Labeled Neurons in Various Brain Structures after Fluoro-Gold Injections into the VTA

		Present study	
	Phillipson (ipsilateral)	Ipsilateral	Contralateral
Forebrain			
Cortex			
Dorsal peduncular	–	+++++	++++
Infralimbic	++	++++	+++
Prelimbic	++	++++	+++
Cingulate	+	+++	+
Agranular insular	+	+++	+
Claustrum	–	+++++	++++
Endopiriform nucleus	–	++++	++
Olfactory tubercle	++	++	++
Nucleus accumbens			
Rostral pole	–	+++++	–
Shell	+++	+++++++	+
Core	–	++	–
Bed nucleus of stria terminalis	+++	++++++	++++
Amygdala	+		
Ant. amygdaloid area		++++	+
Medial nucleus		++++	+
Central nucleus		+	–
Substantia innominata	+++		
Ventral pallidum		+++++++	+++++
Sublenticular subst. innominata		++++	+++
Septum			
Lateral, dorsal part	–	+++	–
Lateral, intermediate part	–	+++++++	+++
Lateral, ventral part	–	++	–
Septofimbrial nucleus	–	+++++	++
Medial/Diagonal band of Broca	+++	++++++	++++
Hypothalamus			
Median preoptic area	–	+++	+++
Medial preoptic area	+	++++++	+++++
Lateral preoptic area	+++	+++++++	+++++
Magnocellular preoptic area	++	++	+
Anterior hypothalamic area	+	+++++	+++
Paraventricular nucleus	–	+++++	+++
Ventromedial hypothal. ncl	–	+	+
Tuber cinereum	–	+++++	++++
Perifornical nucleus	–	++++	+
Lateral hypothalamic area	+++	++++++	++++
Postdorsal hypothalamus	+		
Dorsal hypothalamic area		+++++	+++
Posterior hypothal. ncl		+++++	++++
Supramammillary nucleus	–	+	+
Zona incerta	+	+++++	++
Fields of Forel	+	–	–
Thalamus/epithalamus			
Parafascicular ncl	+	++	+
Paraventricular ncl	–	+++	++
Medial habenula	+++	++++++	++++++
Lateral habenula	++	+++++++	+++++++
Midbrain			
Superior colliculus	++	+++++	+++
Periaqueductal gray	+	++++++	+++++
Substantia nigra	++		
Pars compacta		++++	+++
Pars reticulata		+++	++
Deep mesencephalic field	–	+++++	+++++
Anterotegmental nucleus	–	++++	++++
Ventral tegmental nucleus	–	++++	–
Dorsal tegmental nucleus	–	++	++
Pons and medulla oblongata			
Oral field of pontine reticular formation	–	+++++	+++++
Dorsal raphe	+++	++++++	++++++
Median raphe	+++	+++++	+++++
Paramedian raphe	–	+++++	++++
Pontine raphe	+	+++	+++
Pedunculopontine nucleus	–	++	++
Laterodorsal tegmental ncl	–	++++	++++
Cuneiform nucleus	++	+++	++
Parabrachial nucleus	+++	++++	+++
Locus ceruleus	+	++++	++++
Principal nucleus nV	++	–	–
Caudal field of pontine reticular formation	+	++++	++++
Lateral reticular field	+	++	++
Intermediate reticular field	–	++	++
Gigantocellular reticular field	–	++	++
Cerebellum			
Dentate nucleus	+++	n.d.	n.d.

+, 1–10; ++, 10–20; +++, 20–50; ++++, 50–100; +++++, 100–500; ++++++, 500–1,000; +++++++, >1,000; n.d., not determined.

Figure 2.5: Regions of the brain providing input to the VTA, as determined by injection of the retrograde tracer Fluro-Gold. This list should give some idea of the sheer diversity of inputs to the dopaminergic nuclei. Table taken from (Geisler and Zahm, 2005).

2007).

But once these neurons fire, how does this signal translate into a) the correct behaviour b) future learning by reinforcement?

2.7.2 Dopamine release dynamics

The first level of complexity in the dopamine reward prediction error signal comes about because the release of dopamine is not completely predictable — the amount of neurotransmitter released is dependent upon many factors including the presence of neurotrophic factors and dopamine precursors (Pothos et al., 1998). A phasic burst of dopamine neurons does not lead to the homogeneous increase in dopamine concentration as one might naively expect.

In the striatum it is estimated that only 7% of corticostriatal synapses receive dopaminergic terminals (Arbuthnott and Wickens, 2007), the remainder are dependent upon spillover and diffusion for dopamine modulation. Although we usually assume that dopamine release is global, it is possible that there may be some topographic segregation in the dopamine nuclei. If this is the case it could lead to regionally specific dopamine release and learning. This is an issue which has not been explored in detail for dopamine, but there is evidence for functional segregation in the serotonergic nuclei (Jacobs and Fornal, 1995).

At this stage, after one phasic burst approximately 10,000 vesicles have been released at each axon terminal (Cragg and Rice, 2004). How this goes on to affect target neurons will now be dependent upon diffusion and re-uptake.

2.7.3 Diffusion and re-uptake

When dopamine is released it must diffuse through the extracellular space to reach its target. Because neurons and glia are tightly packed, the space is highly tortuous, so transmitters diffuse more slowly than they would in free space (Cragg and Rice, 2004). It has been proposed by Cragg and Rice (2004) that dopamine diffusion can be approximated according the following equation

$$C(r,t) = \frac{UC_f}{\alpha(4D^*t\pi)^{\frac{3}{2}}}\exp(\frac{-r^2}{4D^*t})\exp(-k't) \tag{2.2}$$

Where $C(r,t)$ is the extra cellular dopamine concentration as a function of distance r, and time t after release. Release is assumed to be instantaneous from a vesicle

with a fill concentration of C_f. Diffusion of dopamine molecules is governed by the local extracellular volume fraction α, and the tortuosity of the extracellular media λ. Tortuosity decreases the diffusion coefficient to D^* ($D^* = D\lambda^2$). The re-uptake of dopamine via DATs and oxidisation via MAOs is incorporated by the uptake constant k'.

While diffusion is taking place to extend the reach of dopamine modulation, there are two processes which will decrease the concentration of dopamine - reuptake to dopamine axons via dopamine transporters (DATs), and breakdown of the neurotransmitter by monoamine-oxidase (MAO). DATs are richly concentrated in the striatum, with lesser numbers in the cortex (Sesack et al., 1998; Lewis et al., 2001). MAO are also found in large numbers in the basal ganglia and hypothalamus, with smaller concentration in the cortex (O'Carroll et al., 1983).

These two processes of diffusion and reuptake/breakdown compete to give the overall concentration of dopamine. In general it has been proposed that DATs limit the temporal, but not spatial extent of dopamine following a phasic burst (Cragg and Rice, 2004). The precise dynamics of this is quite complex, so to understand it in more detail, computational models have been constructed of the process. Prior to beginning the computational modelling described later in this thesis I conducted preliminary work to verify the results reported by Cragg and Rice (2004). I constructed a model of diffusion and re-uptake in the striatum to quantify how transient and local dopamine concentration peaks are relative to the background dopamine concentration that results from the pacemaker activity of dopamine neurons. The aim of this was to provide some quantitative figures of how dopamine concentration might look in vivo given a particular spike train. This data could then be combined with in vitro electrophysiological data to reconstruct the effects of a particular dopaminergic spike train upon downstream neurons.

The model I constructed consisted of a 3D section of simulated tissue, complete with dopamine terminals distributed throughout the space according to the empirical values quoted by Cragg and Rice (2004). An example of the dopamine terminal distribution can be seen in Figure 2.6. These dopamine terminals released dopamine in accordance with the known spiking dynamics. Within the 3D section I calculated the resulting dopamine concentration in a 2D plane. The simulation consisted of 100ms tonic pacemaking activity, followed by a simulated reward (40ms spiking at 80Hz), and then a return to pacemaking activity. The mean concentration in the 2D plane during the simulation is shown in Figure 2.7, and the concentrations in the plane during

tonic and phasic firing are shown in Figure 2.8 and 2.9. The equations and parameters used in the simulations are included in Appendix D.

In agreement with the original study I found that when modelled in 3D, phasic dopamine release results in global, transient (70ms) increase in dopamine concentration to around 7μM. This rise in concentration takes around 20ms, and decays to baseline in 30ms following the offset of the phasic firing. This compares with a relatively constant background concentration of around 0.3uM.

The conclusion of these models is that during a phasic burst the whole of the striatum is awash with a high level of dopamine irrespective of a synapse's distance from the dopamine axon terminals. For an illustration of this see the average dopamine concentration in Figures 2.8 and 2.9 — the few peaks that occur are very short lived. In summary, the peak in dopamine concentration is diffuse in space, but selective in time. There are many components in this model which are simplified or missing (for example the transmitter pool dynamics), but this simple model does give us a preliminary idea of what might be going on. Also it is not clear how these results will compare to the prefrontal cortex where DATs and MAOs are distributed more sparsely.

However, using this model we are able to gain an foothold into how dopamine neuron firing translates into global dopamine concentration. But before we can understand how this translates into the modulation of neurons we must investigate the dynamics of dopamine binding.

2.7.4 Binding dynamics

For dopamine release to have an effect upon neurons it must first bind to receptors embedded in the neuron's membrane. There are 5 distinct receptors that have been identified for dopamine, all of which fall into two groups D1 type, or D2 type (Seeman and Vantol, 1994). The D1 and D2 receptors are the most numerous (hence the names of the groups), and the receptors are present in all regions of the brain where dopamine is released. Dopamine receptors are metabotropic rather than ionotropic, so the effects of binding occur upon the timescales of the intracellular cascades involved (hundreds of ms), rather than the very fast effects that can occur via ionotropic receptors (ms).

The distribution of receptors varies from area to area, with some area such as the striatum showing particularly high concentrations, see Figure 2.10.

As can be seen in Figure 2.10, the relative ratio of D1 and D2 type receptors varies in different regions, and it is thought that the differing ratio of receptors will have func-

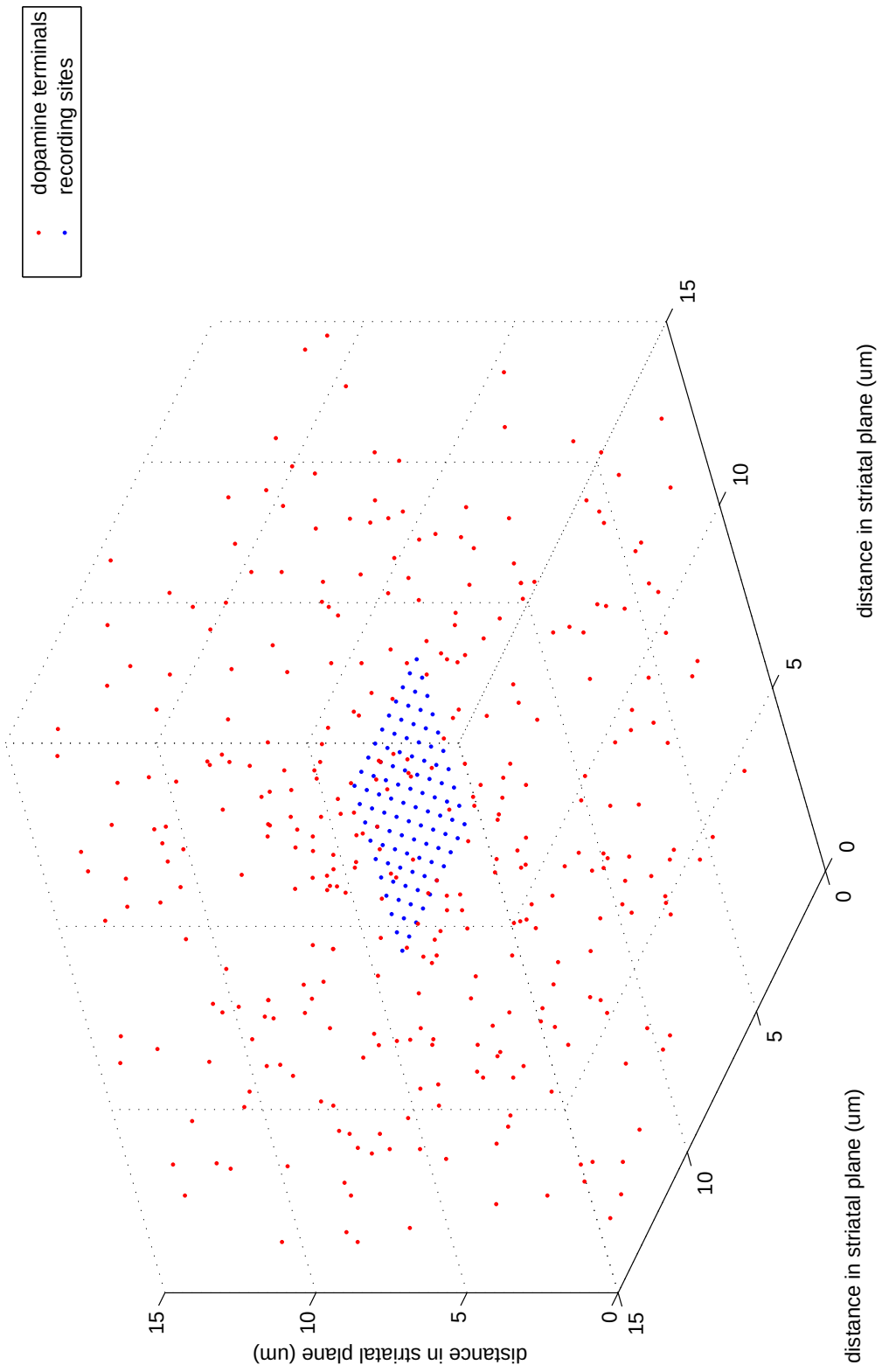

Figure 2.6: The distribution of dopamine axon terminals in the 3D space. Note the 2D recording plane located within the space.

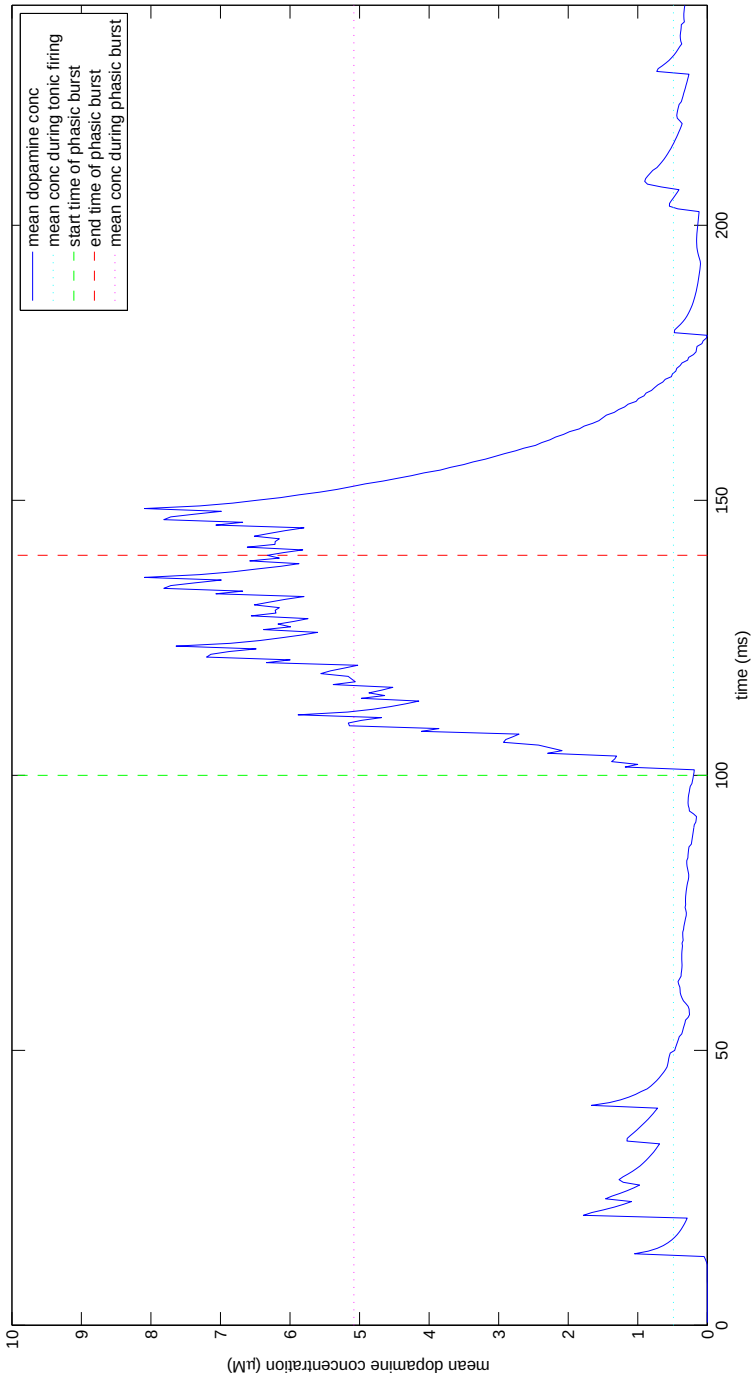

Figure 2.7: The timecourse of mean dopamine concentration in a simulation of dopamine diffusion and reuptake in the striatum. The dopamine concentration values are sampled from a 2D plane of simulated striatal tissue. In this trial dopamine neurons display tonic background firing until the onset of a phasic burst at 100ms. Dopamine release quickly increases throughout the striatum during the phasic burst, and decays to resting concentration 30ms after the phasic burst stops.

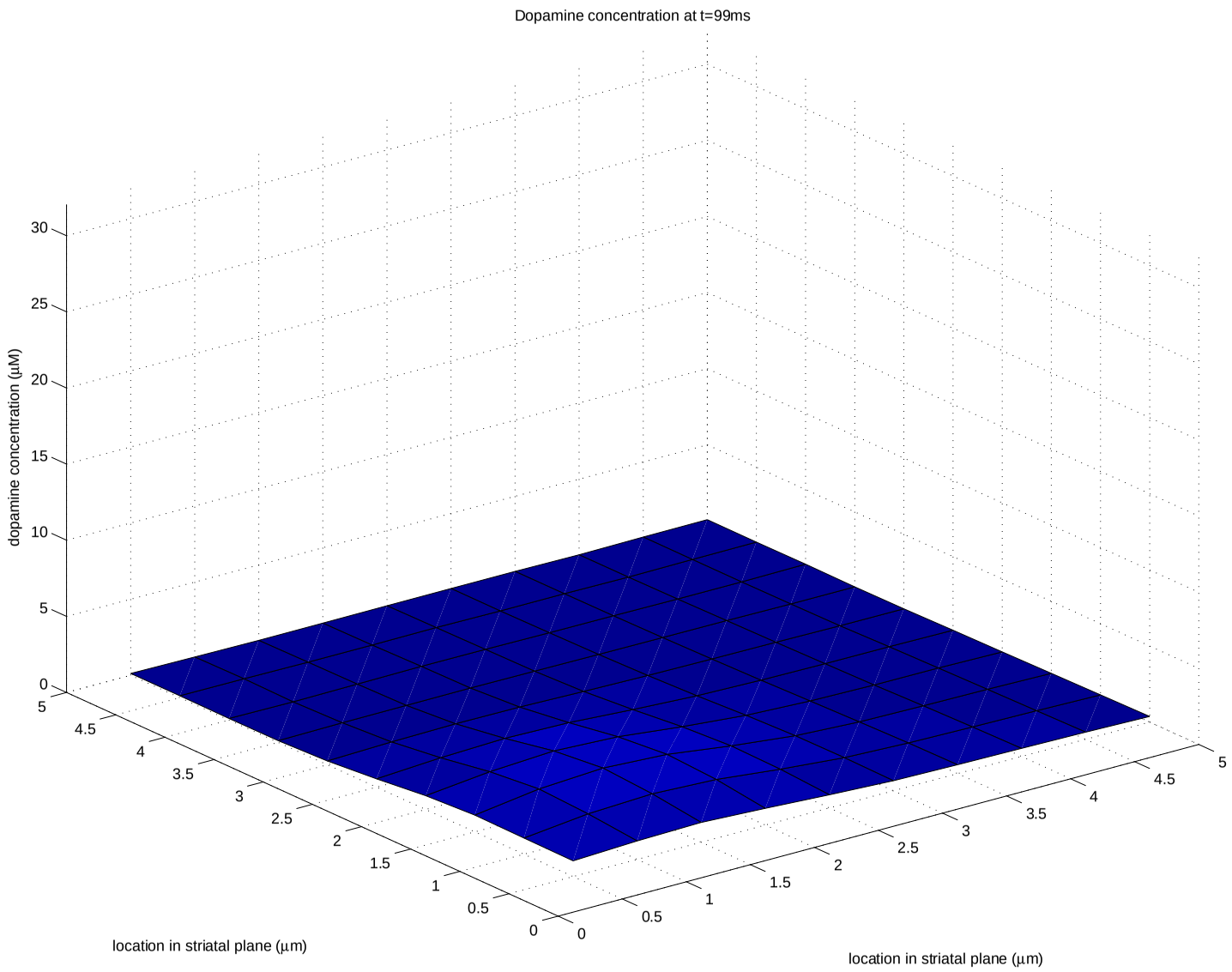

Figure 2.8: A 3D rendering of the simulated dopamine concentration in a 2D plane of striatal tissue. The x and y axes show distance in the plane in μm, whilst the z axis shows dopamine concentration in moles. This graph shows the concentration in the 2D plane at t=99ms in Figure 2.7, when the dopamine neurons are displaying tonic pacemaking activity.

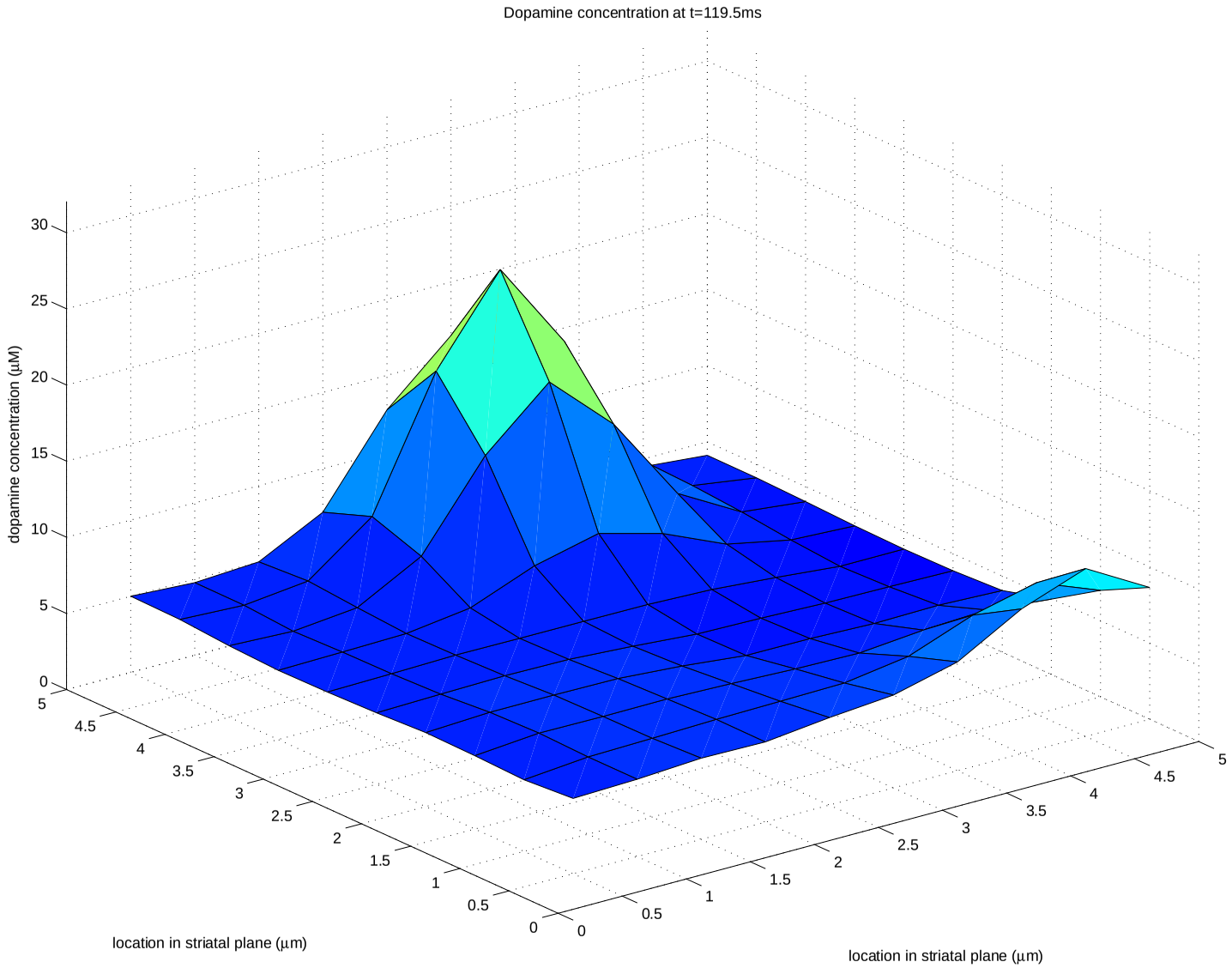

Figure 2.9: A 3D rendering of the simulated dopamine concentration in a 2D plane of striatal tissue. The x and y axes show distance in the plane in μm, whilst the z axis shows dopamine concentration in moles. This graph shows the concentration in the 2D plane at t=119.5ms in Figure 2.7, when the dopamine neurons are displaying phasic spiking. Although there are some regional peaks in concentration for short periods, the main effect of the phasic spiking is to raise dopamine concentration throughout the striatum.

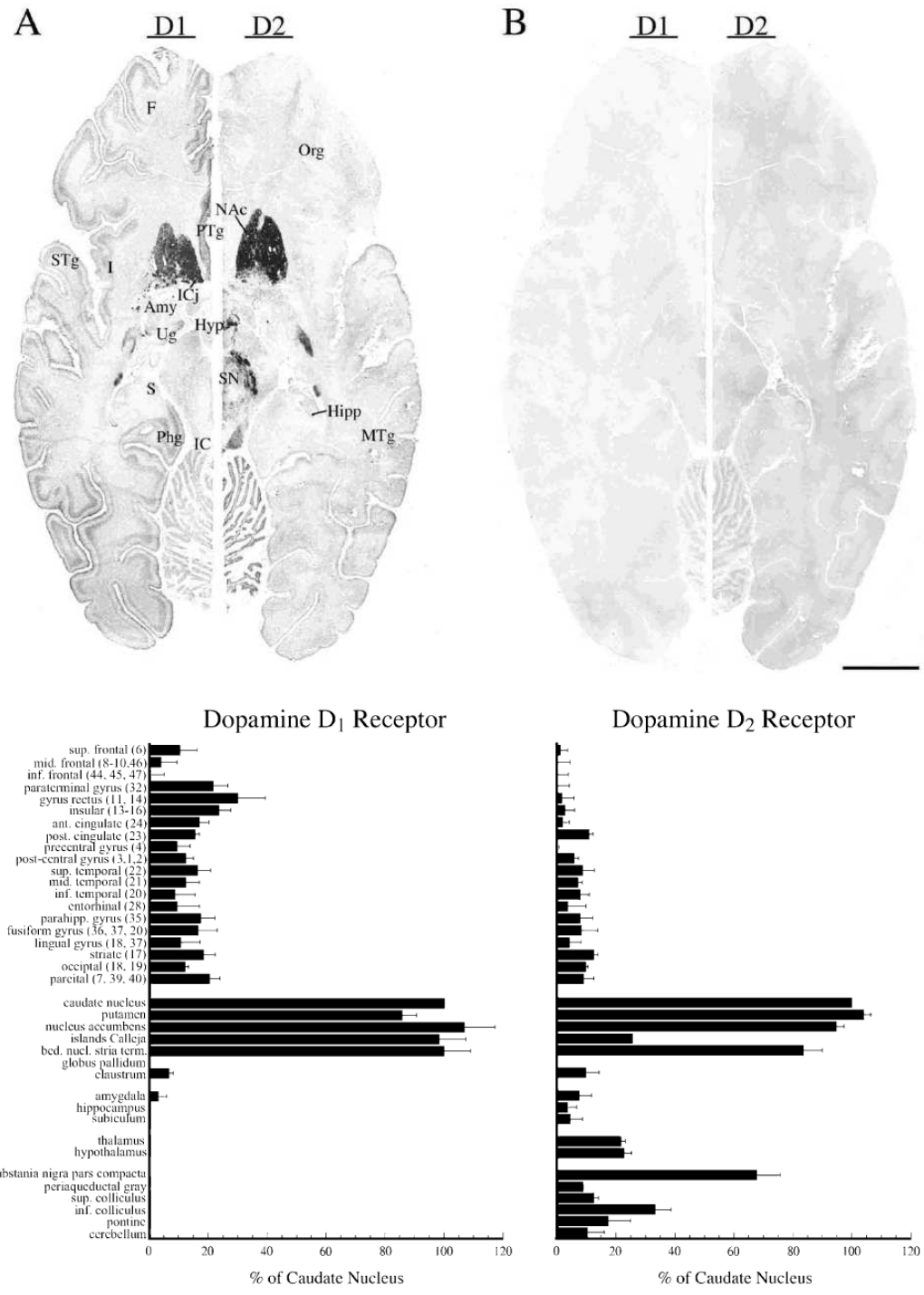

Figure 2.10: D1 and D2 receptor mRNA expression levels in the human brain determined using sense (A) and antisense (B) riboprobes. Taken from (Hurd et al., 2001).

tional consequences (Wang et al., 2004; Mehta et al., 2004). The precise cellular location of the receptors in relation to synapses is not known, but it is thought that they are situated both synaptically (Paspalas and Goldman-Rakic, 2005) and extrasynaptically (Smiley et al., 1994). There is some debate as to whether D1 and D2 type receptors are colocalised upon the same neurons. The evidence so far suggests that somewhere between 20-50% of medium spiny neurons in the striatum colocalise D1 and D2 receptors (Surmeier et al., 1996; Tepper and Plenz, 2006). The rest of the medium spiny neurons appear to act as indepedent pathways, with D1 expressing MSNs projecting to the internal pallidum and substantia nigra pars reticula, whilst D2 expressing neurons project to the external pallidum (Tepper and Plenz, 2006). The low rate of colocalisation may go some way to explaining the heterogeneous response that is observed with in vitro preparations of cortical neurons under dopamine modulation. An example of this can be seen in the in vitro results shown in Figure 4.8 in chapter 4.

Even after the complexities of diffusion and receptor distributions, another layer of complexity is introduced by the varying binding affinities of the different receptor types. D2 receptors have a stronger affinity for dopamine than D1 receptors (Richfield et al., 1989), and so it is likely that these receptors will bind at differing rates during the tonic pacemaking and phasic firing of dopamine neurons. It has been suggested that dopamine that escapes the synaptic cleft will exert its effect mostly though high-affinity D2 receptors (Schultz, 1998) - these kind of assumptions are ripe for testing with a computational model.

In addition to binding dynamics there are also unbinding dynamics. A bound receptor cannot be stimulated, and even after it is unbound, the G-coupled protein must be reset if the neuron is to become sensitive to dopamine again. Therefore, if dopamine release occurs when the receptors are already bound, we would not expect there to be any change to the behaviour.

2.7.5 Intracellular cascades

Once dopamine has bound to a receptor, the G-coupled protein becomes uncoupled, triggering a cascade of intracellular reactions that may result in changes to the intrinsic excitability of the neuron, the synaptic dynamics, or somatic gene expression. This happens at different rates for each receptor, and as we pass further along the cascade it becomes increasing difficult to tease apart the specific effect of each individual dopamine receptor.

D1 type receptors act by stimulating adenyl cyclase and increasing cAMP. D2 receptors on the other hand inhibit adenyl cyclase and cAMP (Tepper and Plenz, 2006). The concentration of cAMP in turn regulates the activity of receptors, ion channels, and transcription factors inside the cell. D1 receptor activation is also capable of mobilising intracellular calcium stores (Neve et al., 2004).

Dopamine also exerts its effect upon neurons via the DARPP-32 cascade, which acts as a converging point for the action of many neuromodulators - dopamine, adenosine, serotonin, GABA, opiates, glutamate, and NO (Svenningsson et al., 2004; Greengard et al., 1999). DARPP-32 is found in all regions with dopamine innervation. The highest concentrations are found in the striatum and substantia nigra pars reticulata, whilst moderate levels are found in the hypothalamus and cortex (Svenningsson et al., 2004). The effects of DARPP-32 are as diverse as its many phosphorylation states, and can manifest themselves in changes to synaptic plasticity or intrinsic excitability of the neuron.

Each of these cascades is highly interconnected, and can be affected by neuromodulators or intracellular calcium dynamics, so we can see it is going to be very difficult to construct a computational model that can incorporate the whole cascade. However, this has not stopped some researchers trying (Lindskog et al., 2006; Fernandez et al., 2006).

2.7.6 Electrophysiological effects

So far we have examined the chain of mechanisms between the phasic firing of dopamine neurons, and the resulting effect upon the neurons involved in and feedback to dopamine neurons. The level of complexity described so far should give the reader a taste for how difficult it will be to describe this process in full with a mechanistic model based upon empirical data.

So far, characterising the electrophysiological effect of dopamine has only been possible in vitro, which makes it difficult to know what additional effect release, diffusion, and binding may have upon the results. It is possible that in the near future techniques like dopamine caging, optogenetics, or voltage sensitive dye-imaging, and pharmacological MRI may be able to help us isolate the effect of dopamine at each step of the process in vivo, but at present we do not have the data available to construct such detailed models.

Typically in vitro experiments are done in a bath of constant dopamine concen-

tration, and so will show quite different behaviour to the in vivo case, where release dynamics, diffusion, and binding will all come into play. However, slice preparations do allow neuroscientists to isolate the particular ion channels that are affected by dopamine modulation, and thereby reconstruct how the spiking behaviour of the neuron is changed by dopamine. This means that for now, any computational model we construct will be largely reliant upon in vitro data for its parameter values.

Much of the electrophysiological work to isolate dopamine's effects has been done in the prefrontal cortex (Seamans and Yang, 2004) or striatum (Gonon, 1997; Calabresi et al., 2007). Here the effects diverge slightly as MSNs in the striatum are inhibitory, whilst pyramidal cells in the cortex are excitatory.

So far in vivo experiments have found that the effects of dopamine are complex. Dopamine has different effects at different timescales (Lavin et al., 2005), and its effects can be inhibitory or excitatory depending upon neuron type and concentration. In the prefrontal cortex the initial effect of dopamine is to increase inhibition due to its effect upon fast-spiking interneurons via D2 receptors (Tseng et al., 2006), but it can also induce upstates and increase evoked spiking in pyramidal cells (Lavin et al., 2005).

In vitro dopamine appears to have a biphasic effect, initially (10-15 minutes) reducing evoked spiking by a D2-dependent mechanism. This is followed by a longer increase in evoked spiking lasting 30 minutes or more, which appears to act via D1 receptors (Seamans and Durstewitz, 2008) It is worth noting that experiments done in vitro to isolate the effects of different receptors or receptor types are done using agonists rather than dopamine itself. These artificial agonists have slightly different binding dynamics than dopamine (Seeman and Vantol, 1994), and this may have some effect upon the results.

In slices the effect of dopamine on synaptic currents can also be investigated. It has been found that dopamine dose-dependently increase NMDA currents via D1 receptors, whilst D2 receptor activation decreases NMDA currents. D1 receptor binding can also decrease presynaptic release probabilities, although this effect may be region specific (Seamans and Durstewitz, 2008).

2.8 Summary

Now we have traced the mechanisms that underlies dopamine modulation, we can now see what is required to verify whether or not the biological data does support or refute

the hypothesis that dopamine is part of a reinforcement learning circuit.

At present there is lots of detailed data about each step of the process, but because of the difficulty of performing in vivo experiments, it has proven impossible to link all this evidence together. My aim in this thesis is to use computational models to combine the evidence where possible into a single model which can be used to test the assumption that the physiology does support a reinforcement learning circuit.

As I have outlined in this brief literature review, there is a huge amount of complexity involved in this process, and so in at least the first iteration we will need to start with a simple model and hope to iteratively introduce more detail. There is a wealth of electrophysiological data on the effect of dopamine in slices, so this may offer us a starting point for an empirically based model. If our model is successful, than there will be the potential to quantify the effects of release, diffusion, and binding in a later model.

Chapter 3

A dopamine-modulated STDP model of Reinforcement Learning

3.1 Novel contributions

This chapter cover the reimplementation and analysis of a model of dopamine modulated reinforcement learning. The model, which is widely cited in the literature is found to solve the distal reward problem in a way that is incompatible with a true model of classical conditioning. The consequences of this are explained in five sections. Potential solutions to these problems are proposed, and the chapter closes with a examination of why these problems were encountered.

3.2 Introduction

My aim in this thesis is to use computational models to examine whether or not dopamine modulation serves to implement a reinforcement learning circuit in the brain.

3.3 Literature review : Existing models of dopamine modulated reinforcement learning

The literature review in chapter 2 discussed the process of dopamine modulation from release to intracellular dynamics, and served to demonstrate that dopamine modulation is highly complex. Rather than start by trying to accomodate all of this complexity immediately, I intend to start with a simple model and work towards one with increasing

biological realism.

There are several models in the literature which attempt to describe how dopamine modulation might underlie reinforcement learning. The earliest from Houk et al. (1994), which made the link between dopamine and actor-critic models of reinforcement learning. This model proposed that eligibility traces such as CAMKII or DARPP-32 in the striatum could bridge the temporal gap between cue and reward, and be reinforcement by the later effects of dopamine. A more formal model by Suri and Schultz (1998) later showed how the actor-critic model could be implemented computationally. Using an implementation of the TD-algorithm they were able to show how phasic firing of dopamine following a reward prediction error might lead to backpropagation of the prediction. An additional model by Contreras-Vidal and Schultz (1999), based upon Adaptive resonance theory offers another route to demonstrating TD-like behaviour based upon the dopamine system.

More recently, a model by Izhikevich (2007) claimed to solve the distal reward problem by dopamine modulated spike-timing dependent plasticity. This model is of particular interest because it proposes microscopic phenomena as the cause of a systems-level behaviour, and claims to successfully implement reinforcement learning. Because I am interested in the mechanics of *how* dopamine modulation could achieve TD-like behaviour I will begin by reimplementing this model with a view to incorporate more biological realism. The aim will be to establish if there is a mechanistic basis for the claim the dopamine is implementing reinforcement learning.

3.4 Dopamine modulation of spike-timing dependent plasticity

The original paper by Izhikevich (2007) claims to solve the distal reward problem through the effect of dopamine on spike-timing dependent plasticity. The paper suggests that the ability of a hypothetical dopamine-modulated STDP mechanism to link rewards to temporally distant cues is enough to allow it to solve a simple classical conditioning task when appropriate anatomical assumptions are made. The model is particularly interesting because it suggests that low-level synaptic mechanisms are the cause of systems level phenomena like reinforcement learning.

If the model is successful then I can use it as a basis for examining the key features of dopamine modulation that are needed to implement reinforcement learning - if I

gradually add more detail regarding the biological effects of dopamine can I verify the original model's assumptions about what dopamine does?

In the interests of clarity I will describe the paper as two models - the first model, which shows in a simple way how a dopamine modulated STDP mechanism can solve the distal reward problem, and the second model, which aims to show how this mechanism can be implemented to solve a classical conditioning task.

3.5 Model 1

3.5.1 Neuron dynamics

The first model is simple and easy to understand, which may be the reason why the paper has provoked so much interest. The neurons are point neurons with no morphological extent, and the membrane potential is modelled using a quadratic integrate and fire equation

$$v_{t+1} = v_t + v_t(0.04v_t + 5) + 140 - u_t + I_{inj}/C + I_{syn}/C \tag{3.1}$$

$$u_{t+1} = u_t + a(0.2v_t - u_t) \tag{3.2}$$

v_{t+1} represents the membrane potential in millivolts at the next timestep given the synaptic input I_{syn}, and an external current injection I_{inj}. Variables u and a are arbitrary values designed to capture the slower membrane dynamics of the neuron. Two sets of these variables are used to represent the two types of neurons being modelled — regular spiking pyramidal cells, and fast spiking interneurons. The parameter values for variables used in this model are included in Appendix A. Code for both the models is included in Appendix B. Details of the neurons, the learning rule, and the results of Izhikevich's simulation are shown graphically in Figure 3.1.

After a spike, neurons are reset to $v = -65$ and $u = u + z$. While the membrane potential of the neuron is measured in millivolts, many of the terms in this model were chosen arbitrarily to reproduce the qualitative dynamics.

3.5.2 Network and connectivity

The network consists of 1000 neurons, each on average synapsing on 100 postsynaptic neurons. The postsynaptic neurons are chosen at random. 80% of the neurons are

excitatory, whilst the other 20% are inhibitory and have synapses that do not show any synaptic plasticity. Each neuron receives a small amount of external stimulation such that the average firing rate of each neuron is 1Hz.

3.5.3 Dopamine and reward

Of the approximately 100,000 synapses in the network, one synapse is chosen at random to be the rewarded synapse. In model 1 an increase in dopamine concentration occurs with a delay of 1-3 seconds any time the neuron postsynaptic to the chosen synapse exhibits a spike within 10ms of the presynaptic neuron spiking.

This dopamine release can be thought of as a delayed reward of the pre-post spike.

When this happens dopamine concentration increases by 0.005uM and decays according to

$$\frac{dd}{dt} = -\frac{d}{\tau_d} \tag{3.3}$$

3.5.4 Synaptic plasticity

All excitatory synapses are subject to an STDP rule such that a pre-before-postsynaptic spike will result in an increase in the eligibility trace for that synapse, whilst post-before-presynaptic spikes will cause a decrease.

The original author proposed that the eligibility trace is "an enzyme important for plasticity", and suggested CAMKII, PKC, or PKA as candidates.

Without any stimulation the eligibility trace for the ith neuron will slowly decay at a rate given by the time constant τ_c

$$\frac{dc_i}{dt} = -\frac{c_i}{\tau_c} + STDP(\triangle t)_i \tag{3.4}$$

The shape of the STDP curve is shown in Figure 3.1b. The area under the long-term depression (LTD) portion of the curve is 1.5 times the area under the long-term potentiation (LTP) portion. This is to ensure that randomly firing synapses tend to see more depression than potentiation, leaving them with lower synaptic weights.

Changes to synaptic strength are governed by

$$\frac{ds_i}{dt} = c_i \times d \tag{3.5}$$

The implication of this learning rule is that if a reward regularly happens shortly after a pre before post event (when the eligibility trace will be high), then the synapse will be potentiated. *If this pre before post event happens with a frequency of 50% or less before a reward is delivered, then the synapse will tend to be depressed* because the LTD region is greater than the LTP region.

When a presynaptic spike occurs, the synaptic current is the synaptic weight multiplied by the capacitance, which is taken to be one — the values here are arbitrary.

$$I_{syn} = sC \tag{3.6}$$

The synaptic weight is bounded between 0 and 4.

3.5.5 Results

The aim of the simulation in this example is to show that dopamine modulation of a simple STDP rule can result in distal cues becoming linked to a reward.

The result reported in the paper is that synapse reaches the maximum allowable value (within a 1 hour period in 42 out of 50 trials).

An example trial is shown in Figure 3.1d.

Figure 3.1d indicates that the model is capable of solving the distal reward problem, but how does it do this? Figure 3.1c shows how pre before postsynaptic activity results in an eligibility trace which slowly decays. If a reward arrives and dopamine concentration increases while the eligibility trace is greater than zero, then the weight will increase, as can be seen in Figure 3.1c. If a pre before postsynaptic spike at a particular synapse is repeatedly rewarded, then an accumulation of these events will lead to the rewarded synapse becoming strong, thereby successfully linking a distant cue with the reward delivery.

The chosen synapse is potentiated more than the others because synaptic change only happens when dopamine is present, and dopamine is present every time there is a pre before postsynaptic spike at the chosen synapse. Other synapses may also demonstrate pre before postsynaptic spikes, but the chances of them doing this while dopamine is present are smaller.

It is important to note that although the eligibility trace is maintained during the delay period, the spikes themselves are not.

Let us assume that when a cue predicts a reward, a chunk of working memory corresponding to that cue stays active during the delay period. This memory trace trig-

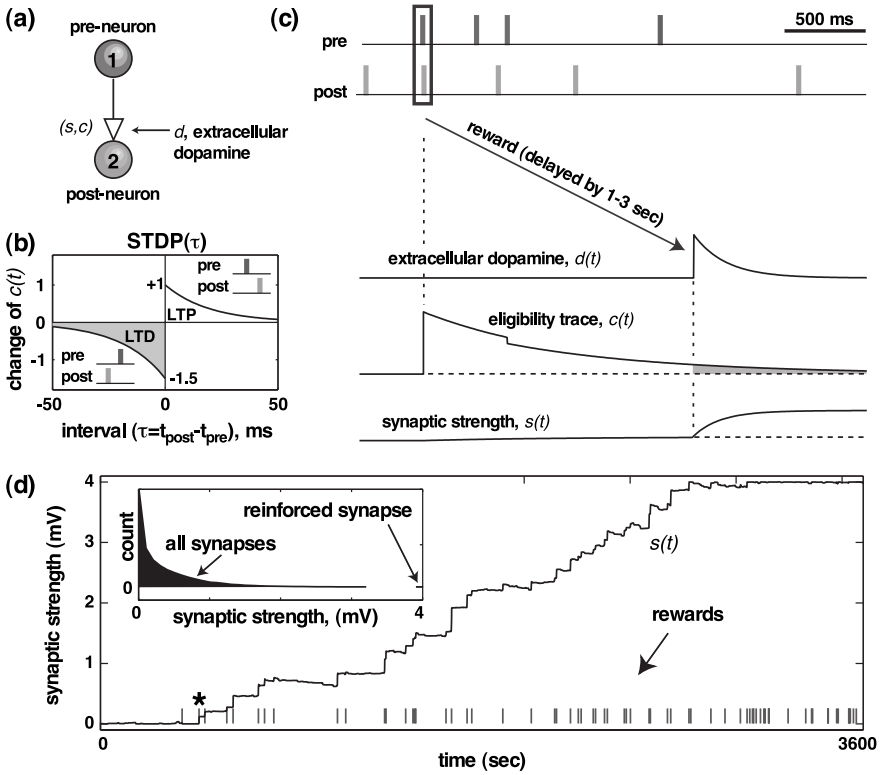

Figure 3.1: The Figure, taken from the original paper shows the model in 4 parts. Part (a) shows schematically that synaptic weights are determined by the interaction of pre and postsynaptic neurons under the influence of a dopamine modulated eligibility trace. Part (b) shows the form of STDP used by this model. LTD is 1.5 times stronger than LTP to encourage random neurons to maintain small synaptic weights. Part (c) shows how the relative timing of presynaptic and postsynaptic spikes interacts to produce the eligibility trace (itself a product of the STDP rule). The increase in dopamine concentration after the reward allows the eligibility trace to have an effect on the synaptic weight. Part (d) shows the results of a typical trial. After many rewarded spikes the chosen synapses tends to increase its weight much higher than other synapses, indicating that the credit assignment problem has been solved.

gered by the cue represents the knowledge that because the cue occurred, a reward is coming. If this is the case, then in this model the memories are stored in the eligibility traces and dopamine concentrations, and not the spikes themselves. This is significant because *if we aim to explain patterns of delay period activity then we ought to choose a model in which the cue trace is represented by activity rather than synaptic eligibility traces*.

Overall the model appears to solve the problem quite well, though it does require a large number of rewards before the synapse is significantly stronger than the others. However, this is just a single synapse, and learning at this synapse cannot be expected to correspond to the learning of high level cues in a conditioning task.

Although this model does successfully link a distal cue and reward, it does not implement backpropagation because the dopamine release does not occur earlier in the trial with each progressive cue-reward pairing. Instead a second model, Model 2, is implemented to show how backpropagation could occur, and how this could be used to solve a classical conditioning task.

3.6 Model 2

The fourth model in the paper, which we will refer to as model 2, builds upon the learning mechanisms described in section 3.5 to implement a model of classical conditioning. In a classical conditioning task an arbitrary stimulus such as a bell is paired with an intrinsically rewarding stimulus (such as food) following a short delay. After multiple pairings of the two stimuli, the animal begins to demonstrate a reward response after the presentation of the arbitrary stimulus alone. The classical example of this is Pavlov's dog — when Pavlov rung a bell on repeated occasions a few seconds before feeding a dog, the dog learned to salivate at the sound of the bell alone. In this case the bell constitutes the conditioned stimulus (CS), and the food the unconditioned stimulus (US).

In Izhikevich's model the author attempts to show how the repeated pairing of a CS and a US can lead to a reward response following a presentation of the CS only. In order to make the analogy between Pavlov's dog and this model clear, we should interpret the firing of VTAp neurons in the model as a correlate of a behavioural reward response, as shown in Pavlov's experiment by the salivation. Comparisons between this model and Pavlov's famous experiment are shown in Table 3.1.

Model 2 functions in the same way as model 1, only in this case some of the 1000

Pavlov's dog	Model 2
food	represented by US neuron firing
bell	represented by CS neuron firing
salivation	triggered by VTAp neuron firing

Table 3.1: An analogy between Model 2 and Pavlov's original conditioning experiments

neurons are separated into groups and labelled as anatomical entities. These anatomical groups are stimulated as a whole when their corresponding stimuli are presented in the task. Stimuli consist of a 20mV injection to each neuron in the chosen group. The anatomical groups in the model are

- The VTAp which represents a cortical area with projections to the VTA. The firing rate of the VTA is assumed to be proportional to the activity of VTAp neurons. Each time a VTAp neuron fires, the dopamine concentration increases by 0.005microM.
- The US which represents the unconditioned, intrinsically rewarding stimulus.
- CS1 and CS2 which represent two different conditioned stimuli

These different anatomical groups are shown in Figure 3.2a. Each anatomical group consists of 100 neurons chosen at random. The US to VTAp synapses are set to the maximal allowable value to represent the strong ability of a primary reward to trigger dopamine release.

3.6.1 Phase 1 : The bell (US)

During the first 100 trials the US alone is presented, which means the 100 US neurons are injected with a superthreshold current. Due to the strong links with the VTAp neurons, this results in an increase in dopamine concentration. Figure 3.2b shows *the response of the VTAp neurons* during trial 100, and the spike raster plot in the lower part of figure 3.2b shows *the response of a typical neuron* in the VTAp group in each of the 100 trials.

3.6.2 Phase 2 : The bell then the food (CS1 → US)

During trials 101-200 the neurons in the CS1 group are stimulated and after a random delay of 1 ± 0.25s the US neurons are stimulated. This is intended to mimic a classical conditioning experiment where a CS and US are repeatedly paired with some short interval. In Figure 3.2c we can see that by the final trial, the VTAp response to CS1 is increased.

3.6.3 Phase 3 : Bell 2, then Bell 1, then the food (CS2 → CS1 → US)

In trials 201-300 the CS2 stimulus is introduced before CS1, and again in Figure 3.2d we can see that the response of the VTAp shifts to CS2.

This behaviour, where the earliest informative cue triggers dopamine release rather than the reward is similar to what is observed in experiments, and has been described by theorists as a backpropagation of reward prediction error. The author uses these results to conclude that the model is a successful implementation of Pavlovian and instrumental conditioning.

3.7 The practicalities of extending upon model 2

The initial aim of this chapter was to extend upon model 2. However, when I tried to reproduce the results, I came across several problems, which I will discuss briefly here.

At first I attempted to reproduce the model myself based upon the description in the paper, but was unable to obtain the same results. I consulted the published code and found that this only included a cleaned up version of the code for model 1, and not the actual code that was used to generate all of the results in the paper. I contacted the author who agreed to email me the code that had been used to generate the data. There were some minor discrepancies between what his code did and what was described in the paper, but the basic result stood. After reading the original code I discovered that I had been unable to reproduce the results due to some ambiguities in how the model was described in the paper. There was also a misunderstanding about what was being shown in one of the diagrams, and I was only able to resolve this after studying the original code.

On further study of the code I discovered that there was a lot more going on than was evident from the original paper. The next section will go into detail on this. I used

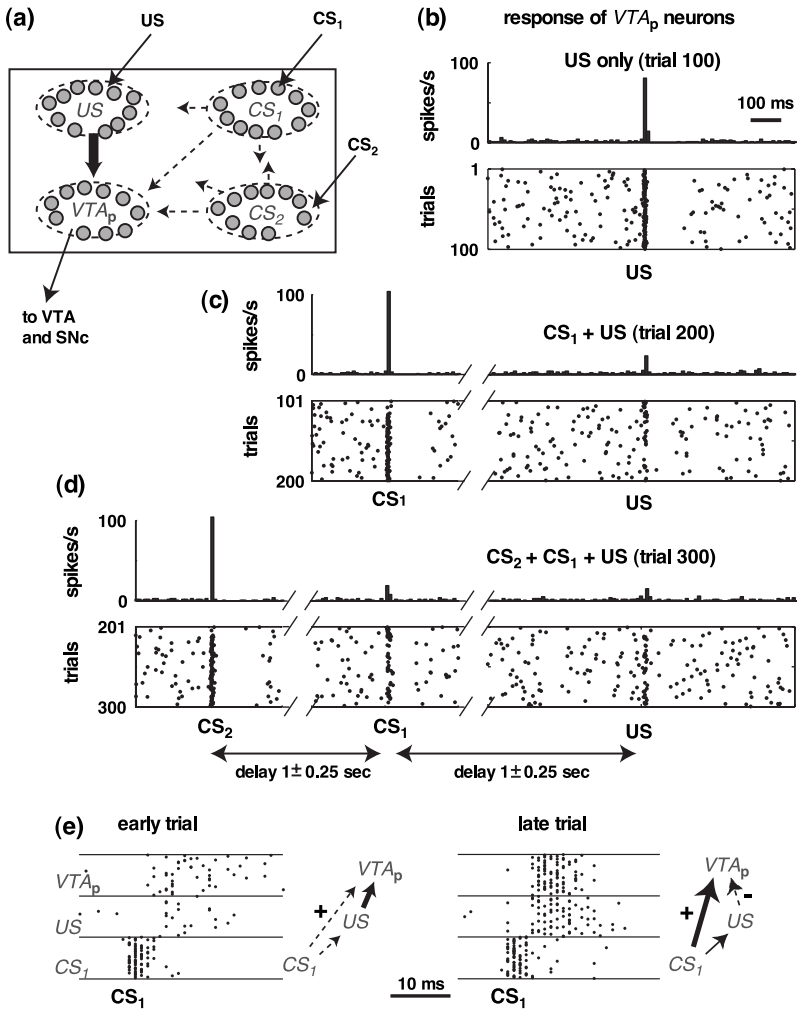

Figure 3.2: The figure, taken from the original paper shows model 2, the results, and an explanation of how these results come about. Part (a) shows the anatomical groupings in the model and how they are interconnected. Parts (b), (c), and (d) show the results for trials 1-100, 101-200, and 201-300. In each set the upper graph shows the total VTAp responses in the last trial, and the lower graph shows the VTAp responses over the whole range of trials. Part (e) attempts to explain how the apparent backpropagation occurs. The arrows suggest that learning of the CS → VTAp synaptic link is bootstrapped by the initially strong CS → US → VTAp link.

a modified version of the code for the analysis in the following sections, and that code is included in Appendix B.

3.8 Problems with the model

As described in section 3.6, the paper claimed that the model was capable of completing a simple classical conditioning task, and indeed it was capable of producing the graphs in Figure 3.2. However, after analysing the original code I discovered that the model displayed certain behaviours that were *incompatible with a true model of classical conditioning.* In addition some of the explanations given in the paper were misleading or incorrect.

The author does describe some limitations of the model in the original paper, but these do not relate to the fundamental problems that were found in my analysis. While I could attempt to address each of the original caveats individually, I feel that the severity of the problem explained in section 3.8.1 renders such a discussion unneccessary.

3.8.1 The US is not necessary for learning the CS

The model will learn anything, irrespective of whether or not it is paired with a US. A set of simulations designed to illustrate this are shown in Figure 3.3.

To highlight how critical a problem this is, we could compare this to the famous Pavlov's dog experiment — it is as if the dog is learning to salivate at the sound of the bell even if it is never paired with the food. As long as stimulus is presented enough times, the dog will learn to salivate when it occurs — clearly if the model does this then something is fundamentally wrong. If the US is not necessary for learning the CS, then the explanation shown in Figure 3.2e is incorrect because the CS $\rightarrow$ US $\rightarrow$ VTAp link is not involved in the process.

3.8.2 Learning is too quick

If we examine the synaptic strength between groups at the end of each trial we might get the impression that learning is happening gradually. However, if we plot what is happening within trials as in Figure 3.4, we will see quite a different story.

The synaptic changes that do occur are not a result of gradual learning at a behavioural timescale, but are due to almost instantaneous changes that happen at synapses when the neurons are artificially stimulated. This means that synaptic weights are

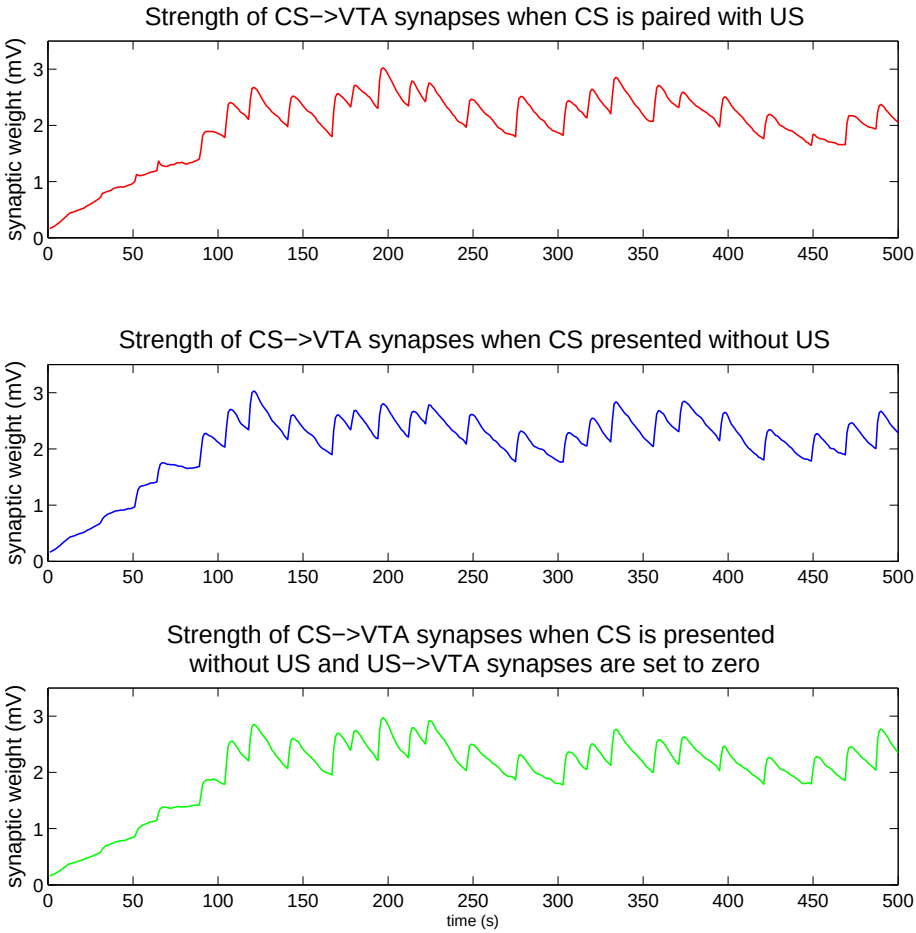

Figure 3.3: The upper graph shows the strength of CS → VTA synapses starting from when the CS is first presented. As we can see the strength of these synapses increases initially and then stabilises. Izhikevich (2007) suggests that this process is aided by the presence of strong US → VTA links. However, if we do not present the US at all (the middle graph), or we do not present the US AND lower US → VTA synapses to zero, the CS is still learned. This indicates that the explanation given in the paper is incorrect — in this model, feeding the dog is not necessary for learning for the dog to learn to salivate at the sound of the bell.

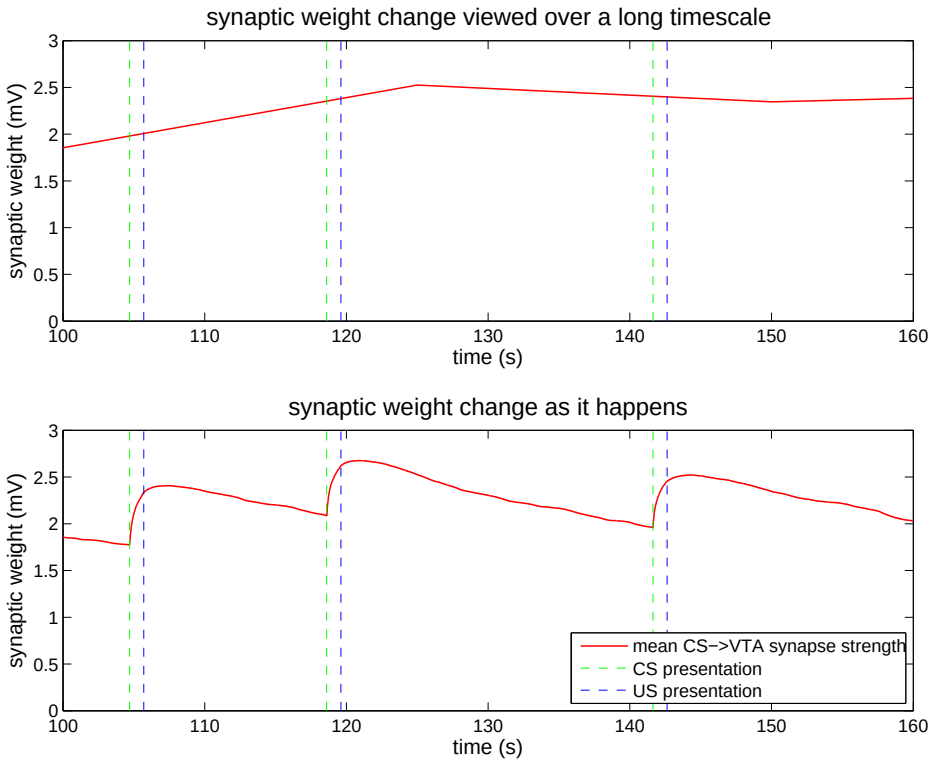

Figure 3.4: The upper graph shows the mean CS → VTA synaptic weight as sampled at one point during each trial. This gives the impression that the synaptic weight changes are taking place gradually over the course of a few trials. However, the lower graph shows the synaptic weight change as it happens (10ms resolution). Here we can see that synaptic weight change is taking place rapidly within trial, and the largest changes occur when the external stimuli are injected into the network.

changing rapidly during the trial. This kind of behaviour is not desirable because we are implicitly assuming that synaptic weights represent the animal's belief about the probability of a reward following a particular cue. Our belief about the probability of a cue signalling a reward is something which changes slowly, over the course of many trials, and ought not to change significantly during a single trial.

When I state that learning is too happening too quickly I am not referring to the number of trials that are required for learning, but the fact that rapid synaptic change is happening *within* trials .

3.8.3 Learning is indiscriminate

Figure 3.2e in the original paper is misleading because it appears to show that the CS comes to trigger dopamine release through a specific CS->US->VTA pathway. In fact the stimulation that occurs when the CS is presented is so strong that *all synapses post-synaptic to CS neurons will see their synapses potentiated*, and not just CS->US synapses (a schematic of this process is shown in Figure 3.6). In fact all synapses from the CS are enhanced, not just CS-US or CS->VTA (see Figure 3.5). In this sense learning is indiscriminate.

3.8.4 US response is depressed by an unrealistic mechanism

In the previous section I demonstrated that the potentiation of responses to the CS is happening for different reasons than is suggested in the paper. But what about the depression of the US response following CS presentation that occurs in phase 2? (shown in Figure 3.2c). If this is not happening due to a backpropagation of a prediction error, then how is it happening?

Figure 3.2c shows that once the CS is presented, the response of the VTA neurons to US stimulation is decreased. In experiments with animals this is thought to happen because the conditioned stimulus, such as the bell in the case of Pavlov's dog, comes to give all the information about whether the reward is due to arrive or not. If the bell is rung then the dog knows with 100% certainty that the food will arrive, and therefore there is no reward prediction error when the reward does arrive — it was expected. But if we look at how this process happens in the network in model 2 we can see that it is not due to a gradual backpropagation of certainty as the dog becomes more sure that the bell predicts the food. In the model this process happens by two mechanisms: one a side effect of how the neurons work, and the other an unintentional consequence

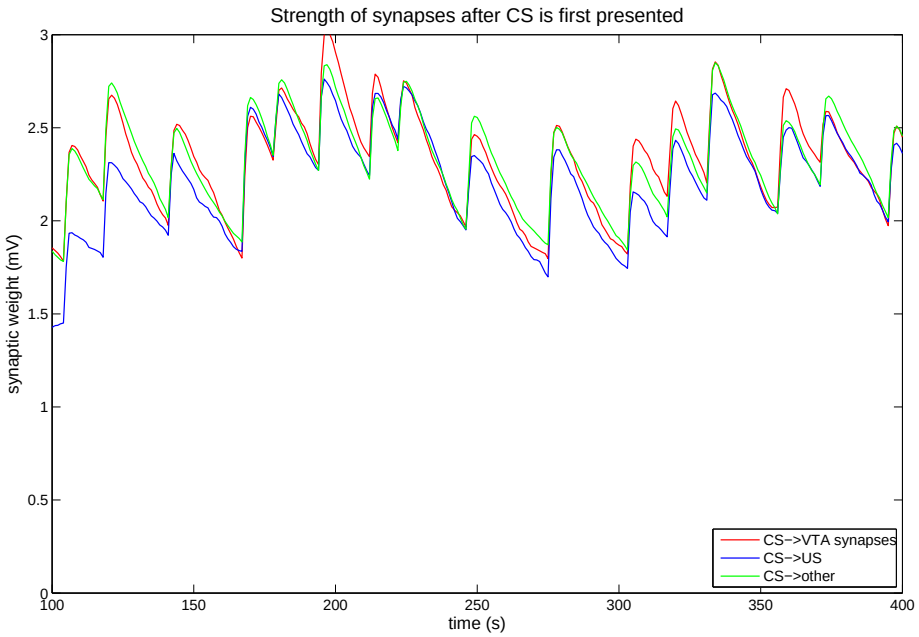

Figure 3.5: The graph shows synaptic weight change over the course of many trials for a selection of synapses postsynaptic to CS neurons. As we can see the synaptic weight changes are very similar, irrespective of their relationship to reward. It appears that when a strong external stimulus is applied, all neurons are potentiated, and not just those connecting to the VTA.

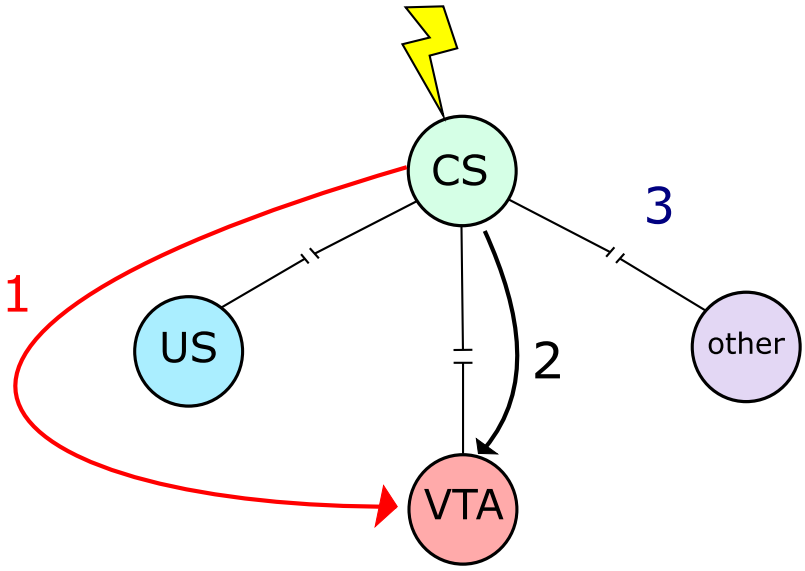

Figure 3.6: In the original paper the author suggests that the increase in synaptic weight of the CS → VTA pathway (labelled in the figure as pathway 2) is initially due to its indirect links to the VTA via the US (pathway 1). However, in Figure 3.5 we can see that all synapses are potentiated at the same rate irrespective of their links to the VTA via the US. Also, Figure 3.3 shows that if the US is not presented at all and the US → VTA synapses are weak, then the synapses on pathway 2 are still potentiated at the same rate. This indicates that the explanation given in the paper is incorrect.

of stimulating the network too hard when the CS is presented. What happens in the network can be seen in Figure 3.7.

To summarise, the US is depressed when the CS is presented because:

1. The US stimulus is less able to cause postsynaptic neurons to spike because so many neurons have been reset following the CS stimulus. This results in a lower dopamine concentration, and a smaller eligibility trace.

2. The feedforward activation of US neurons that occurs after a CS stimulation causes many US neurons to spike just after dopamine has been released. Due to the way the STDP rule has been set up, this causes strong depression of US synapses after a CS stimulation

Although both these mechanisms produced the desired result — that is a depression of the US response as the CS is learned, neither of these mechanisms is supported by empirical evidence, particularly the second. The first mechanism is sensitive to the time delay between the CS and the US, and is only likely to be effective when there is a very short gap between the cue and the reward — longer than one second and the effect is likely to disappear. The second mechanism suggests that mere presentation of a CS causes unconditioned rewards to become less rewarding. This would be problematic because over an animal's lifetime it is exposed to a continuous stream of stimuli, and only some (the reward predicting ones) ought to have the effect of reducing the ability of the unconditioned stimulus to elicit reward.

3.8.5 Dopamine release is happening in an artificial way

A fundamental problem I encountered with this model is that learning is occurring at the time of the cue due to the *cue related dopamine release that is caused by the artificial stimulation of the network*. During the first pairing of the CS and US, the presentation of the CS causes greater dopamine release than the delivery of the reward itself, and this does not fit any known experimental data (see VTA spikes in trial 101 of Figure 3.7). Although particularly salient cue stimuli are known to cause the release of dopamine in experiments, it is not of the same magnitude as the phasic dopamine release that occurs when there is an unexpected reward.

This CS-related dopamine release allows learning to occur immediately following the cue presentation, which is why the bell can be learned as a predictor of reward, even when no reward is presented. If dopamine is released after the cue, then there is

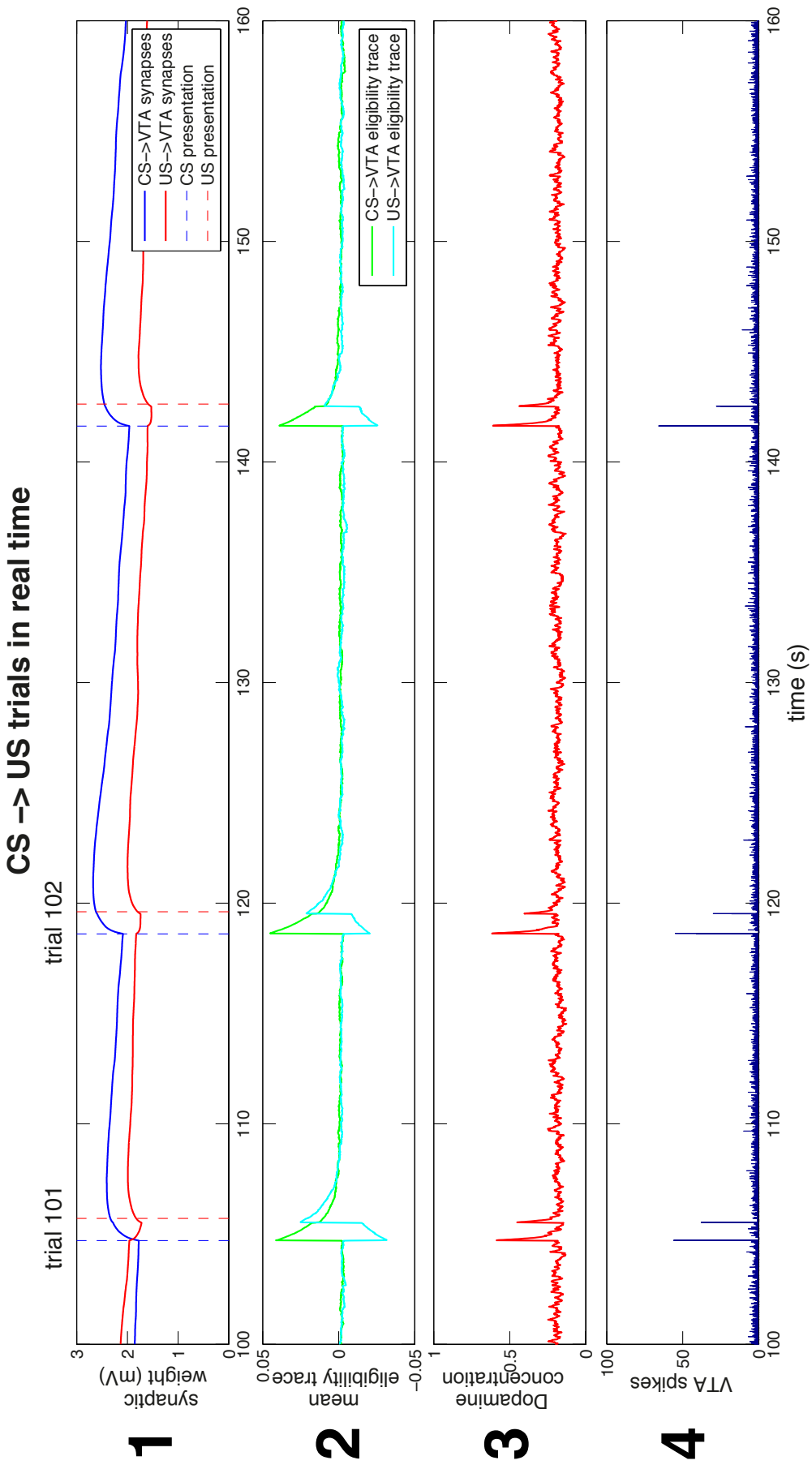

Figure 3.7: During trial 101 US → VTA are stronger than CS → VTA synapses (trial 101, graph 1), yet stimulation of the US results in fewer VTA spikes than stimulation of the CS (trial 101, graph 4). This is because after the CS stimulation, many US neurons are still within their "refractory" period and do not fire. This results in a lower eligibility trace in US neurons due to fewer postsynaptic spikes following presynaptic spikes (trial 101 graph 2), and less dopamine release (trial 101, graph 3). The second part of the depression mechanism kicks in in the next trial. After trial 101, CS → US synapses have been potentiated (trial 102, graph 1), and so when the CS is stimulated during trial 102, this causes both US and VTA neurons to spike immediately. Any US neurons that are slower to spike, and spike in the following timestep are likely to have their synapses depressed as they will spike at a time when the postsynaptic neurons have already spiked, and Δt is small and negative. This will result in a negative eligibility trace (trial 102, graph 2) at a time when dopamine concentration is high. This explains why US → VTA synapses are depressed at the time that the CS is presented. Clearly, this immediate depression due to feedforward activation in the network is not the same process as genuine backpropagation.

no incentive for the stimuli to stay active, and therefore this mechanism cannot explain patterns of activity during the delay period.

3.8.6 Reward prediction error is not backpropagated

Although the model does appear to solve the distal reward problem, it does not do this by propagating certainty from the time of the reward back to the cue. *Learning in this model occurs immediately at the time of cue presentation* due to feedforward activation following a stimulus. It is not a result of bridging the temporal gap between the cue and reward. Izhikevich (2007) alludes to this in his comment that CS-VTA learning occurs 'on a compressed timescale'. Although this may appear to be a subtle point, it does have significant consequences.

If learning occurs through backpropagation then this may lead to persistent activity during the delay period, which, as we explained in chapter 2, is known to happen in working memory and reinforcement learning experiments. If learning occurs due to feedforward activation, as it does in this model, then an additional mechanism will be required to explain the persistent activity during the delay period.

Another well known feature of neurons shaped by reinforcement learning tasks is that they begin to show ramping activity that conveys information about the timing of various actions during and at the end of the delay period. Again, this is something that might occur if learning is backpropagated across the delay period, but will need an additional mechanism if the synaptic changes are due to feedforward activation.

3.9 Potential Solutions

This section covers potential solutions to the problems outlined in the previous section.

3.9.1 Novelty

The CS can be learned without the US because the external stimulation of the CS is strong enough to release dopamine, without a reward (US) ever being presented.

The significance of using a CS is that it is a stimulus that is not intrinsically rewarding in itself, so the fact that it can be learned as a predictor of reward represents a problem. One solution to this problem could be to decrease the original strength of the external stimulations by including a mechanism to reduce the ability of cues to induce

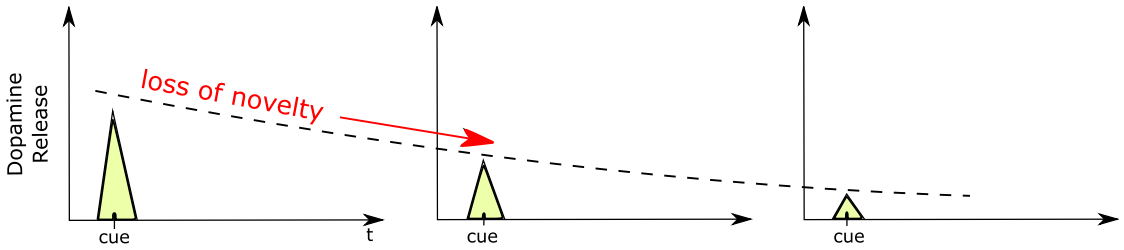

Figure 3.8: A novelty mechanism would mean that a new stimuli would elicit a diminishing release of dopamine with each presentation

dopamine release as they become less novel. An example of what this would mean in terms of dopamine release is shown in Figure 3.8.

It is known from experiments that particularly salient or strong stimuli that do not in themselves lead to reward, can often cause the release of dopamine — for example switching on a light, or introducing a new object to the animal's surroundings. It is thought that this dopamine release is due to the novelty of the stimulus, rather than any link between the stimulus and reward.

So, rather than weaken the stimulations in the model and thereby decrease the level of dopamine release following a stimulus (which might stop the other results working), we could introduce a mechanism which would decrease the salience of the cue stimuli as they become less novel

It is possible, that if set up the right way this would prevent CS being learned as predictors of reward without the US present

3.9.2 Implementing backpropagation of reward prediction error

Unfortunately the other problems in the model are not easy to solve. The fact that the model is solving the distal reward problem through feedforward activation, rather than backpropagation is a fundamental problem — it isn't something quantitatively wrong which can be fixed by tweaking parameters, it is a qualitative problem with how the model solves the problem. I outlined in section 2.4.2 the reasons why backpropagation is a suitable mechanism for solving the distal reward problem — it can potentially explain many phenomena in one fell swoop. If we were committed to using the current learning mechanism for model 2 we would have to introduce additional mechanisms to solve these other phenomena, which decreases the parsimony of the model.

Although I have shown how most of the learning in model 2 occurs due to feedforward activation, the original mechanism described in model 1 is still present in model

2, and is doing backpropagation - it is just that the addition of strong external stimuli in model 2 swamps out this original mechanism.

So, is there a way that the strength of the external stimuli could be reduced to make this backpropagating mechanism more prominent? Possibly, but it is questionable as to whether work to do this would be of much value. As mentioned in the analysis of model 1, the back propagation across the delay period that occurs in the first model is by means of the eligibility trace, and is not due to persistent delay period activity of neurons.

As the time of dopamine release propagates backwards, it is the patterns during the delay period that reliably led to reward will be set in stone — be they eligibility traces or neural activity. If we are looking for a mechanism that might explain the patterns of neural activity in the delay period that are observed in experiments, then the memory trace of the cue ought be stored in neural activity itself, and not eligibility traces.

3.10 Summary of findings

1. The CS can be learned as a predictor of a reward without presentation of the US, which is in contradiction with a basic result in classical conditioning.

2. Learning in the model is too quick and occurs over a few seconds during the course of a trial. As a result the synaptic weights cannot be said to be representing reward probabilities, as our beliefs regarding reward probabilities should not change within the course of a trial.

3. Learning is indiscriminate — the CS becomes associated with all neurons in the network, not just the US.

4. Reward prediction error is not propagated backwards from the reward presentation to the time of the first predictive cue. This means that the interesting delay period dynamics observed in experiments - persistent or ramping activity — must be explained by an additional mechanism.

5. Because the reward prediction error is not being propagated backwards we are reliant on a unrealistic depression mechanism. In this model US $\rightarrow$ VTA synapses are depressed as a quirk of the stimulation process.

3.11 Discussion

My initial intention when choosing this model was to build upon its success in solving the distal reward problem. However, in the process of replicating the result I discovered that the model was not working in a way that was compatible with a model of classical conditioning. After analysing the problems in the model I concluded that it would be easier to start a new model from scratch than to add fixes to the existing model. While the effect of dopamine upon synapses is undoubtedly important, I am not convinced that a model of reinforcement learning should be implemented at the synaptic level. Following my experience with this model I would opt to implement a model of reinforcement learning at the level of populations of neurons — a strategy I will follow in chapter 5.

I will separate my discussion of this model into two sections, one dealing with issues regarding the theoretical basis of the model, and another dealing with the practical issues of relating it to the data.

3.11.1 Theoretical Issues

3.11.1.1 Why base the model on synapses?

I have argued in section 1.4 that when we seek to explain some phenomena we should base it upon correlates at an appropriate spatial and temporal scale. This model is unusual in the way that it seeks to explain systems level behaviour (reinforcement learning) in terms of microscopic correlates changing at a much faster timescale (synaptic plasticity). So it seems an obvious question to ask — If the author was intent on explaining a systems level effect, why did he choose to begin with a model of synaptic dynamics? Is this choice of mechanism a purely rational one, or does it reflect a priori assumptions that are often made in computational neuroscience?

I have argued in chapter 2 that Donald Hebb's ideas have been highly influential in theoretical neuroscience, and as a result many of our computational models follow him in ascribing learning to changes in synaptic plasticity. Since Hebb published his ideas in 1949, we have discovered many other substrates of persistent change in the brain, yet many computational models of learning and memory still insist on implementing abstract models of synaptic process occurring on the timescale of milliseconds to explain behavioural processes occurring at the timescale of seconds and minutes. For these reasons I would argue that it is better to use a simpler model of dopamine's effect

on memory at a systems level, rather than to invoke a dopamine modulated STDP rule.

3.11.1.2 Why use STDP?

So far I have argued that some of the choices made in constructing the model reflect theoretical assumptions in the field, but I believe there are also sociological factors behind the choice of the model's components. It is important to note that the author chose not just to implement a model of synaptic plasticity, but rather opted to implement STDP. STDP is currently a popular model of synaptic plasticity, but there is no evidence to suggest it is particularly relevant to the reinforcment learning process.

3.11.1.3 Why is STDP popular?

To take another step back in our analysis, we might wonder why is it that STDP is popular with theorists in the first place?

In a sense, some timing-dependence of plasticity is inevitable. Spikes cause time-dependent changes at the synapse, and so it is clear that the degree of plasticity will somehow be dependent upon the relative timing of spikes.

When STDP was first proposed and found empirically, it was in the form whereby presynaptic spikes before postsynaptic spikes cause potentiation, and postsynaptic spikes before presynaptic spikes cause depression - this is often informally referred to as "vanilla" STDP. The vanilla form of STDP was particularly appealing to theorists (and perhaps also the reason why experimentalists went looking for it) because it implied that presynapses that spike at the right time to *cause* postsynaptic spikes ought to be "rewarded" (note the overlap in vocabulary here). This kind of rule is appealing, because it would appear to *reinforce* presynaptic spike patterns that lead to postsynaptic spikes — an important property if we believe that information in the brain is coded in the temporal patterns of spikes, which is a popular view in computational neuroscience (see section 1.2.4). The vanilla STDP rule is set up in such a way as to maximally potentiate causal spikes, and depress spikes which are not likely to be causal. So in a sense, the vanilla STDP rule was favoured over other acausal rules

1. Because it is likely to support and lead to the formation of temporal codes of spikes

2. Because it is a restatement of the basic scientific assumption of causality and purpose. Spikes which are not causal appear as noise, while spikes which are casual appear to have a purpose (they cause their effect).

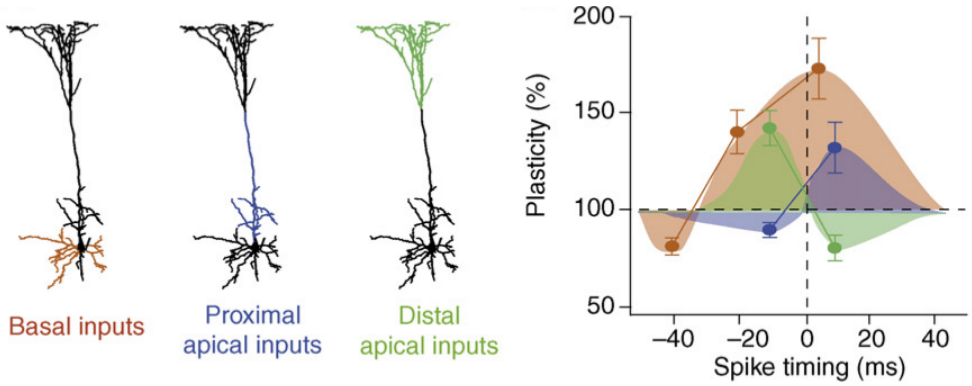

Figure 3.9: A graph taken from (Kampa et al., 2007) showing the diversity of STDP curves than can be found just by measuring in different locations in the dendritic tree. Spike timing shown on the x-axis refers to the time of the post-synaptic spike minus the time of the pre-synaptic spike.

These properties make the vanilla STDP rule both appealing on on theoretical grounds, and also useful in situations where a causally sensitive learning mechanism is required to preserve causal relations in the data. But what is happening here? Are we solving a problem using empirically discovered and motivated mechanisms, or are we merely looking for data that fits our assumptions (vanilla STDP data), extracting generalisations from that data (the vanilla STDP rule), and reapplying these "mechanisms" to new problems (reinforcement learning) in the hope of showing that our original assumptions can also be found to be upheld in other domains? Is this process one of objective analysis, or are we oblivious to our own biases?

Perhaps this line of argument may be controversial, but what cannot be argued is that the assumptions of the model in question are undersupported by the data. In particular it is known that STDP can occur in many different flavours (see Figure 3.9), but it is the one that reinforces casuality that was first predicted and first observed. If there are many different possible STDP curves, how can the author's choice to use the vanilla STDP curve be justified, particularly when the whole result is dependent upon choosing this particular curve?

It is easy to draw the conclusion that the vanilla STDP rule was chosen because it had properties that are desirable for a model where causal relationships are to be reinforced — if the learning rule is rewarding causal spikes at the level of synapses, then this is likely to be better at reinforcing causal relationships at the systems level than an acausal learning rule would be.

This rationale is entirely hypothetical and may not reflect the process by which

the original author came to develop the model. However, I choose to highlight it here because I believe that some of the problems I have had with this model reflect problems that will be found with a whole class of models that are common in computational neuroscience.

This way of working, where mechanisms with the right kind of properties are assembled to solve a problem is a suitable way to work in engineering, but not in science. By using this kind of engineering approach it is often quite easy to "solve" the problem, much in the same way as this model solves the distal reward problem. But solving the problem with a collection of hand-picked mechanism does not guarantee that the model is an accurate representation of how the brain solves the problem.

If this kind of engineering approach to modelling is commonplace, how do we avoid falling into this trap with future models? One approach would be to work with the literature, and construct models using only mechanisms and anatomy which are known to have strong support from the experimental data. Another approach would go even further — to attempt to make the model as much of a one-to-one model of the physiology as is practically possible. With these kinds of models every mechanism and parameter setting would correspond to something which is practically measurable by an experimentalist. Both of these methods have their advantages and shortcomings, and will be examined in more detail in the following chapters.

3.11.2 Practical Issues

Aside from the theoretical issues with how the model was constructed, there are also practical issues with relating to experimental data. The model makes some assumptions about the effect of dopamine modulation on synaptic plasticity that are not supported by any empirical data. While the the ability of the model to link distal cues and rewards is interesting, the lack of empirical evidence does make it difficult to support the model as a mechanistic description of how the brain implements a reinforcement learning circuit.

The mechanisms used in the model are undersupported by the empirical data, so the only conclusion we can draw is that these mechanisms were chosen to ensure that the model would solve the problem. Although solving the problem does make the result seem more impressive (see the title used for the paper), the result is only interesting to neuroscience if it is *the* mechanism by which the brain solves the problem.

In this model, as is the case with many models in computational neuroscience it is

easy to get the impression that the theory came before the data. The problem with this is that theoretical mechanisms are rarely easy to relate to empirical data, because by nature they are abstracted from empirical data. If we want to assess the validity of this model our key question is — *"does the solution presented in this model match what we see when we perform experiments?"*. This question can be answered most easily if components in the model correspond to measurables.

3.12 Conclusion

Although part 1 of the model does link distal cues and rewards, the classical conditioning model 2 in the second half of the paper fails to reproduce some of the basic features of classical conditioning. While some problems with the model may be fixable by adding new mechanisms, others seem to indicate that the way in which the model learns must be fundamentally changed. The model makes some claims regarding the nature of dopamine modulation which are not supported by empirical data, and as a result it is difficult to support or refute the model with empirical data. This suggests that future modelling work should be more empirically driven in order to be verifiable by experiment.

Chapter 4

The effect of dopamine modulation on a model of synaptic plasticity

4.1 Novel contribution

A computational model of the effect of dopamine modulation on synaptic plasticity is constructed. The chapter introduces a simple model prefrontal cortex pyramidal cell, and uses empirical data to parameterise a model of how this neuron will behave under dopamine modulation. The computational model is used to assess the relative contribution that dopamine modulation of firing rate and synaptic conductance has on synaptic plasticity. The results indicate that the synaptic effects of dopamine are more significant than the effects upon intrinsic excitability, particularly in the case of D1 receptor modulation. The results from this computational model are compared with recent experimental results.

The chapter ends with analysis of lessons learned from developing more empirically driven computational models.

4.2 Empirical Modelling

In the previous chapter I looked at a systems level model of how dopamine modulated STDP might implement a reinforcement learning circuit in the brain. At the end of the chapter I concluded that the engineering-style approach that had been used to construct the model made it difficult to support or refute the model with the empirical data because it wasn't closely linked to empirical data.

One way of guaranteeing that our models are tied to the biology and can be verified

by experiment is to base our models purely upon empirical data and see what patterns we observe. It is these patterns in the data which we ought to raise to the status of causal mechanisms, and not mechanisms such as STDP which are preferred on a priori grounds.

The aim of this chapter is to explore this approach to modelling within the bounds of the problem at hand — the hypothesis that dopamine signals a reward prediction error and thereby forms part of a reinforcement learning circuit in the brain. Can we put experimental observations into a computational model, hit the "RUN" button and expect to generate results that support or contradict the hypothesis? Or is this too simplistic a view?

4.2.1 Investigating reinforcement learning with an empirical model

In the previous chapter, the model described in (Izhikevich, 2007) made some basic assumptions about the effect of dopamine on synaptic plasticity, and went on to show that these simple mechanisms could have far reaching effects. However, none of the assumptions that were made were directly supported by empirical data — for example the assumption that STDP occurs only in the "vanilla" STDP form, or the assumption that dopamine acts as an across-the-board enhancer of plasticity.

The case for dopamine's involvement in reinforcement learning is based upon the observation that dopamine neurons exhibit phasic spikes in a similar manner as had previously been predicted by abstract models of reinforcement learning. Beyond this correlation between the reward prediction error of temporal differences algorithm and the activity of dopamine neurons, we have no clear models of what phasic dopamine release does, and *how* it contributes to the broader task of learning to repeat reinforced behaviour. What does it mean to say that dopamine acts as a reward prediction error 'signal'? What does this abstract concept of a signal mean when we are talking about the brain? How can a neuromodulator "signal" something?

If my aim in this thesis is to develop a systems level model of the role of dopamine modulation in a reinforcement learning circuit, then I will also need a high level description of the effect of dopamine on plasticity that is empirically justifiable. In order to do this I aim to develop an empirically based model of the effect of dopamine modulation on synaptic plasticity, and use the results from this model to feed into the systems level model.

In this chapter I will take the known electrophysiological effects of dopamine on

prefrontal cortex pyramidal cells, and plug this data into an existing model of synaptic plasticity.

4.2.2 Setting up a model

The aim of the model will be two-fold

1. To try out a more empirically motivated approach to modelling
2. To examine the effects of dopamine modulation on the synaptic plasticity of prefrontal cortex pyramidal cells, and to produce observations that can be used in a systems level model of dopamine modulated learning.

There is a large amount of electrophysiological work that has been done to characterise the effects of dopamine on pyramidal cells, both in terms of changes to the intrinsic excitability of the neurons, and also changes to the synaptic properties of the neurons. *If the hypothesis is correct* (that dopamine forms part of a reinforcement learning circuit), *then the electrophysiological effects of dopamine ought to lead to an improved ability to predict reward*.

Saying that "dopamine release leads to an improved ability to predict reward" is quite an easy statement to make, but it is practically a very difficult hypothesis to test. How does one detect whether or not this is happening, either in an experiment, or in a computational model? How do we know if small changes at synapses will lead to an improved ability to predict reward without already making some assumptions about how the brain predicts reward? If we intend to measure reward prediction, then we must have some metric and an idea of what to measure, and this in itself constitutes a theory.

4.2.3 Signs of successful reinforcement learning

If I aim to assess whether or not dopamine modulation of synaptic plasticity is supporting reinforcement learning I must have a clear idea of how this happens, and therefore be able to determine whether or not the results support or contradict the hypothesis.

1. One mechanism which has been proposed as a key mechanism in learning to predict reward is STDP. It has been suggested (Abbott and Blum, 1996) that spike timing dependent plasticity offers a means for neurons to learn temporally integrated sequences of events — a property which is critical for successful

reinforcement learning. This property is critical because at its heart, reinforcement learning is a problem of learning to link temporally separated events. If STDP allows us to do this, then as shown in the last chapter, a dopamine modulated STDP mechanism may support reinforcement learning and allow back-propagation to occur. If dopamine modulation does form part of a reinforcement learning circuit, evidence that dopamine modulation causes an enhancement in the efficacy of pre before post synaptic spikes would offer some support to the hypothesis. However, if dopamine modulation results in an enhancement of anti-hebbian spike timing dependent plasticity, (an increased efficacy of post before pre synaptic activity) this would suggest that at the single synapse level, dopamine modulation tends to act against mechanisms that might aid reinforcement learning. This might be used as evidence to suggest that dopamine is not solely acting as a reward prediction signal, and that therefore the reinforcement learning model applied directly to neurotransmitters like dopamine is too simplistic.

2. Another very simple but important observation is that behavioural reward is a very strong learning mechanism, and that if dopamine is proposed to be acting as a signal of unexpected reward, then *an increase in dopamine concentration ought to result in an increased rate of synaptic change*. Such a significant increase in synaptic plasticity should be observable even using simplistic artificial plasticity protocols. This effect ought to be substantial, and potentially of a greater significance than changes to the timing-dependence of plasticity.

4.3 Literature review

It has been hypothesised that dopamine provides a reward prediction error signal as part of a reinforcement learning circuit in the brain. This hypothesis is based upon the observation that dopamine neurons fire phasically when animals encounter a difference between predicted and actual reward (for a review, see (Schultz, 1998)). However, it has not been explained how the neurons that are modulated by dopamine use the signal to learn the appropriate conditioned response. One possibility is that dopamine leads to learning of the conditioned response through persistent changes at synapses.

4.3.1 The effect of dopamine on synaptic plasticity: Empirical work

Experimental work suggests that the primary factor which causes synapses to change their the efficacy is calcium influx. The calcium influx that follows a postsynaptic action potential is capable of triggering internal processes which change the sensitivity of the synapse to further spikes. The degree of calcium influx that follows a postsynaptic spike is partially dependent on whether or not a back-propagating action potential propagates from the soma to the dendrites and synapse (Spruston et al., 1995). Schiller et al. (1998) have shown that calcium influx through NMDA receptor is amplified following postsynaptic action potentials. These two results suggest that it is the interaction of the back-progating action potential and the NMDA receptor that causes the amplification in calcium influx. If I intend to build a model of the effect of dopamine on synaptic plasticity, then it is crucial that the simulated effects of dopamine modulate this calcium influx mechanism.

Changes in synaptic efficacy can be quantified according to many factors - the most commonly manipulated varibles being the rate of presynaptic stimulation, the cooperativity of input, and the relative timing of pre and post-synaptic spikes. The degree by which each of these factors effects plasticity outcomes has been studied in depth by Sjöström et al. (2001) who showed that plasticity in the cortex is dependent upon both frequency and spike-timing, and that the two factors interact to determine the change in synaptic efficacy.

Alongside this work on plasticity in control situations, there has also been experimental work done to investigate plasticity in the presence of dopamine. Otani et al. (1998) found that the presence of dopamine facilitated LTD in slice preparations, and can also facilitate LTP if the slice is "primed" with a stimulation before application of high-frequency stimuli in the presence of dopamine (Blond, 2002). This priming with dopamine was observed when dopamine was applied to the slice 30 minutes before a paired application of dopamine and tetanic stimulation. This suggests that dopamine triggers long-lasting changes in the prefrontal cortes that can manifest themselves in synaptic plasticity long after the transient increase in concentration.The latency of these effects is on timescale of minutes, suggesting that dopamine effects plasticity both in the long-term,and in the short term via the calcium influx mechanisms described above.

Work in the prefrontal cortex has shown that D1 receptor activation in the prefrontal cortex facilitates LTP, and LTP in the prefrontal requires NMDA receptor activation

Gurden et al. (2000). Together these results suggest that dopamine facilitates LTP via the effects of D1 receptors on NMDA channels. At the behavioural level, Baldwin et al. (2002) have shown that operant conditioning in rat requires D1 and NMDA activation in the prefrontal cortex.

Since the modelling work in this chapter was completed there has been additional experimental work to determine the effect of dopmine modulation on synaptic plasticity. Pawlak and Kerr (2008) found that dopamine receptor activation was required to observe spike-timing dependent plasticity at cortico striatal synapses. Also, in the striatum, Shen et al. (2008) showed that D1 and D2 receptor activation has bidirectional, timing-dependent effects on plasticity. In the prefrontal cortex Xu and Yao (2010) found that dopamine enabled the induction of timing-dependent LTP through D1 and D2 receptor activation. D1 receptors act via a cAMP-PKA signalling mechanism, whilst D2 receptors facilitate LTP through their modulation of GABAergic circuits.

4.3.2 The effect of dopamine on synaptic plasticity: Computational modelling

A number of computational models have set out to address this problem, and have tried to create mechanistic models of how dopamine-dependent plasticity during rewarded tasks can lead to predictions of future rewards. Many of these models have assumed that dopamine modulation leads to reinforcement learning behaviour through it's effect upon STDP (Thivierge et al., 2007), (Florian, 2007), (Izhikevich, 2007). Other models, such as that by Nakano et al. (2010) have instead examined the effect of dopamine on plasticity via it's effects on intracellular cascades within the synapse.

Although these models have attempted to characterise the effect of dopamine on STDP or synaptic mechanisms, there is also more general electrophyisological data that could be used to quantify the effects of dopamine modulation on plasticity. Seamans and Durstewitz (2008) reviews data which could be used to quantify the effect of dopamine upon synaptic conductances. In vitro work has also begun to tease apart the effect that dopamine modulation has on activity over long and short timescales - Lavin et al. (2005) report that in the prefrontal cortex, the response to a phasic burst of dopamine neurons involves first inhibition, and then later a gradual potentiation of responses that lasts for tens of minutes.

Some models have attempted to use this data to capture the effect of dopamine on on neural excitability. Gruber et al. (2003) and Durstewitz et al. (2000b) have devel-

oped models of the effect of dopamine modulation upon firing dynamics of neurons in the striatum, and the prefrontal cortex respectively.

Both of these types of models have produced interesting results, but none of the models has looked at the effect of dopamine upon *BOTH the excitability of the neurons AND the plasticity processes at the synapses.* As a result it is not clear which of these two mechanisms has the most significant effect on learning. If we are to develop an empirical model of dopamine modulation which can be used as a basis for a systems-level model, we need to assess the effects of both these pathways to plasticity.

4.3.3 Summary

In summary, empirical work on the effect of dopamine upon the excitability of neurons has not yet been integrated with existing systems level models of synaptic plasticity. In order to fill this gap in our knowledge I will use this chapter to develop an empirical model that attempts to quantify the relative contribution to synaptic plasticity of dopamine modulation of neural excitability, and dopamine modulation of synaptic conductances.

4.4 Methods

The experimental hypothesis was tested by constructing a computational model of prefrontal cortex neurons and running simulated plasticity protocols under varying levels of simulated dopamine modulation. In practise this meant that the simulation included a model post synaptic neuron which was stimulated by a model pre-synapse in accordance with an in vitro plasticity protocol. Each plasticity protocol was run three times to simulate

1. a bath solution of 50μm SKF38393 (a D1 receptor agonist)
2. a bath solution of 10μm Quinpirole (a D2 receptor agonist)
3. a control (ie. no dopamine modulation)

In order to characterise the effect of dopamine modulation on synaptic plasticity, an existing calcium based model of plasticity was used to calculate the change in synaptic efficacy due to calcium concentration in the post synaptic neuron.

In this section I describe the model in four parts:

1. The neuron model
2. The plasticity model
3. The model of dopamine modulation
4. The simulated plasticity protocols

4.4.1 The neuron model

The computational model neuron consists of a minimal set of equations required to reproduce the spiking dynamics of a prefrontal cortex pyramidal cells. Using the methods described by Izhikevich (2006) a three-conductance model was developed and parameterised based upon in vitro data to show a similar spike shape to a regular spiking prefrontal cortex pyramidal cell (4.1). Although the three equations are not intended to model specific ion channels, they are loosely based upon sodium and potassium conductances. The model was implemented in Matlab using Matlab's variable timestep ode solver. All of the parameters used in this model are included in Appendix C.

There do exist other, more detailed models of prefrontal pyramidal cells, such as the one described by Durstewitz et al. (2000a). This model implements Hodgkin-Huxley like conductances (based upon in vitro recordings) to recreate the spiking behaviour of in vitro cells. However, the minimal model was chosen because it produces qualitatively similar spike shape whilst being orders of magnitude faster to simulate than the morphologically detailed neurons used by Durstewitz et al. (2000a).

I chose to use a simple model neuron because I intended to reuse the results in larger, network simulations which would run much more quickly if based upon a computationally inexpensive neuron model. The neural model was deemed accurate enough on the basis that the membrane potential trace closely fits traces from in vitro recordings of prefrontal cortex pyramidal cells.

Crucially the model neuron fires from a plateau as is observed in in vitro recordings. The choice of this simple neural model had knock-on effects which are described in section 4.4.3. The equations specifying the model are:

$$\frac{dV}{dt} = (I_{inj} - I_{leak} - I_{Na} - I_K - I_M - I_{AMPA} - I_{NMDA})/C \tag{4.1}$$

$$I_{leak} = g_L(V - E_L) \tag{4.2}$$

$$I_{Na} = g_{Na} m_\infty q (V - E_{Na}) \tag{4.3}$$

$$I_K = g_K n (V - E_K) \tag{4.4}$$

$$I_M = g_M h (V - E_M) \tag{4.5}$$

$$\frac{dq}{dt} = \frac{q_\infty - q}{\tau_{Na}} \tag{4.6}$$

$$\frac{dn}{dt} = \frac{n_\infty - n}{\tau_K} \tag{4.7}$$

$$\frac{dh}{dt} = \frac{h_\infty - h}{\tau_M} \tag{4.8}$$

Here the different currents I of the neuron are represented by the subscripts $leak$,Na,K, and M for the leak, sodium, potassium, and M conductances respectively. V represents the membrane potential, and E the various reversal potentials. The terms q,n, and h represent gating variables which are given by the following equations

$$m_\infty = (1 + \exp^{(V_{Nahalfmax} - V)/k_{Na}})^{-1} \tag{4.9}$$

$$q_\infty = (1 + \exp^{(V_{NaInhalfmax} - V)/k_{NaIn}})^{-1} \tag{4.10}$$

$$n_\infty = (1 + \exp^{(V_{Khalfmax} - V)/k_K})^{-1} \tag{4.11}$$

$$h_\infty = (1 + \exp^{(V_{Mhalfmax} - V)/k_M})^{-1} \tag{4.12}$$

The values of these conductances were set by parameter fitting so that both the membrane potential trace and the f-I curve of the neuron closely matched the experimental data for the control condition. The parameter search was done using an iterative interval bisection method, and fits were assessed according to the sum of the squared error.

The method for parameter fitting for the conductances g_L, g_{Na}, g_K, and g_M was the same as was used for fitting the dopamine modulated f-I curves, and is described in more detail in section 4.4.3. This level of parameter fitting was considered reasonable as for a given input the model neuron produces a spike train similar to that observed experimentally. In our case we are interested in the spiking behaviour of the neuron, and not necessarily the specific ion channel dynamics which can be computationally expensive to simulate.

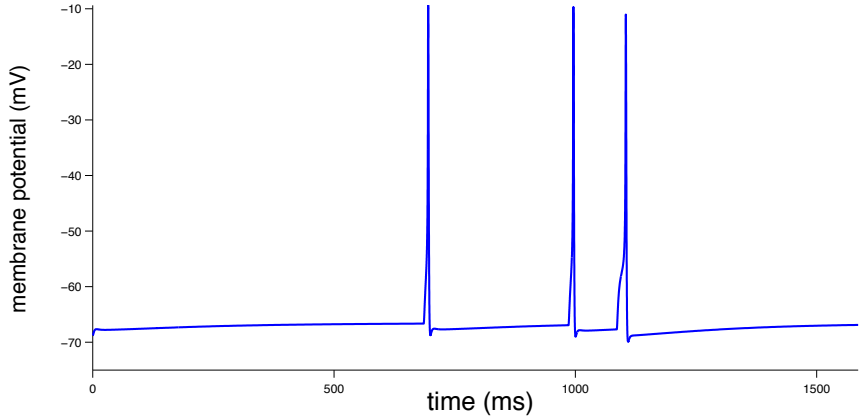

Figure 4.1: An example membrane potential trace for the model neuron. Although the neuron consists of only 3 equations, the spike shape qualitatively matches that of a real prefrontal pyramidal cell in that the neurons spike from a plateau.

At the synapse, AMPA, NMDA, and GABA conductances were simulated as the sum of two exponentials following Durstewitz et al. (2000a) (Figures 4.2, and 4.3). The calcium concentration at each synapse was the product of the NMDA conductance and the conductance of c_∞, which is intended to represent the effects of a back-propagating action potential on voltage-dependent calcium channels. The dynamics of c_∞, and the effect on calcium influx can be seen in Figures 4.5 and 4.4).

$$I_{AMPA=}g_{AMPA_{max}}[\tau\tau_2/(\tau_2-\tau_1)]*[\exp(-t/\tau_2)-\exp(-t/\tau_1)] \tag{4.13}$$

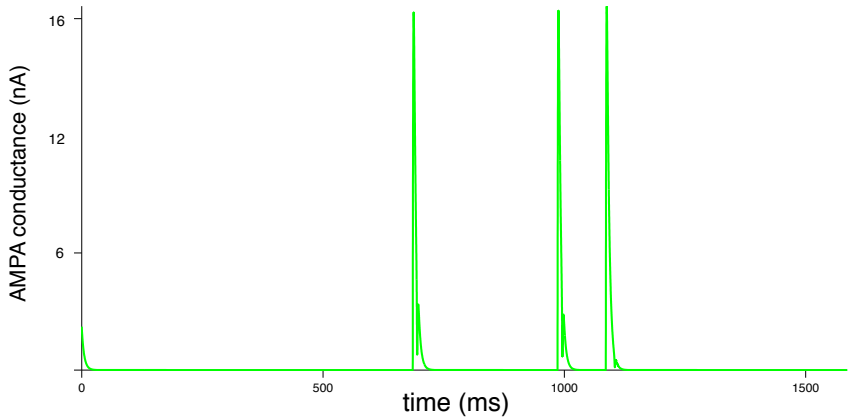

Figure 4.2: The current through the AMPA channel is given by the equation above. The graph shows the AMPA current corresponding to the neuron voltage trace in Figure 4.1. The AMPA current is characterised by a fast onset and offset, here controlled by the two time constants $\tau_1$and τ_2.

$$I_{NMDA=}g_{NMDA_{max}}s*[\tau\tau_2/(\tau_2-\tau_1)]*[\exp(-t/\tau_2)-\exp(-t/\tau_1)]*V_1 \quad (4.14)$$

$$s = (1+0.33\exp(-0.0625V_1)^{-1} \quad (4.15)$$

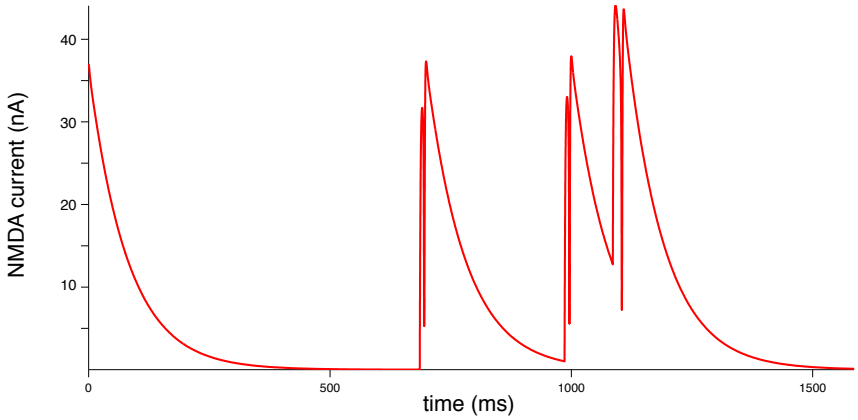

Figure 4.3: The NMDA current is similar to the AMPA channel, but with a slower offset and some voltage dependence. The slow offset of the NMDA conductance means that calcium can continue to enter the synapse long after the spike. The voltage dependence of the NMDA channel makes it sensitive to coincidences between pre and post synaptic spikes — an ideal quality for a spike-timing dependent plasticity mechanism. The voltage dependence of the channel comes from the membrane potential term V_1 and the contribution of the voltage-dependent Magnesium block s, whose equation is also given above. The conductance shows some artefacts where it decreases immediately after a spike. This is due to the fact that the model neuron hyperpolarises after a spike - an unfortunate consequence of the using just three terms to model the neuron (see Figure 4.1). Although the degree of hyperpolarisation can be changed by altering the parameters, the qualitative shape of the curve cannot be altered without adding terms to the neuron model. However, this may not be neccessary - the artefact is so short-lived that it does not have a significant impact on the total conductance through the channel after a spike.

$$I_{Ca} = g_{Ca} I_{NMDA} * c_{\infty} \tag{4.16}$$

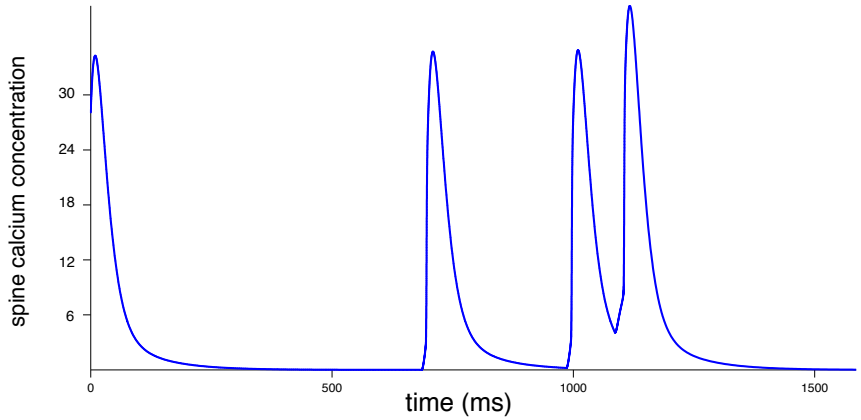

Figure 4.4: The model assumes that the main source of calcium influx at the synapse is through the NMDA channel and voltage-dependent calcium channels (represented by c_{∞}). c_{∞} is graphed in Figure 4.5. The contributions of the NMDA channel and c_{∞} are multiplicative to represent the non-linear relationship between voltage and calcium influx, and also to reflect the contribution of calcium-induced calcium release from intracellular stores. This multiplication between the fast $c_{\infty}$and the slow I_{NMDA} is crucial, because the slow offset of the NMDA channel makes it too slow to explain the tight millisecond timing-dependence required for a symmetrical STDP curve. The model has been parameterised to work with the plasticity model, so the units here are effectively arbitrary.

$$c_\infty = 0.0005 + 0.0095 * (1 + \exp((-20 - V_1)/2))^{-1} \quad (4.17)$$

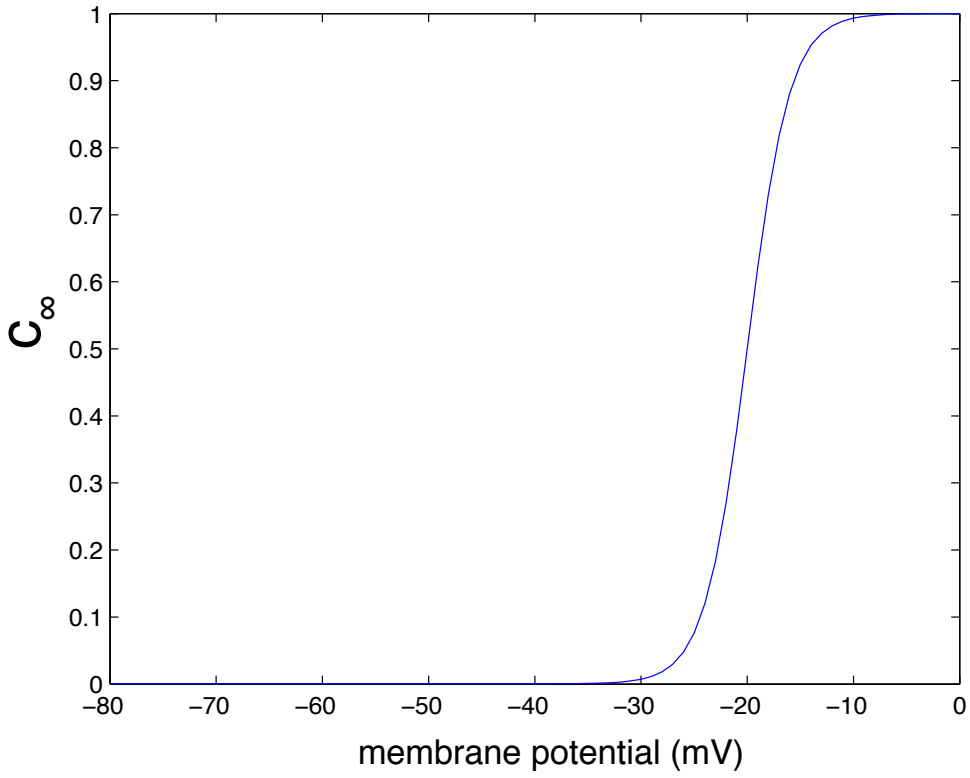

Figure 4.5: A graph showing the conductance through c_∞ as a function of the membrane voltage. The parameter values were chosen so that the channel opens as the neuron depolarises. The result of this is that the peak calcium influx will be during a spike.

The product of the NMDA current and c_∞ rather than the sum was used so as to reproduce the non-linear increase in calcium influx found experimentally during a postsynaptic spike. The channel is added multiplicatively rather than additively not just to produce a non-linear increase in calcium influx, but also to cause the calcium influx to shut off more quickly after a spike than would a NMDA channel. This can be seen in the difference in decay rates between Figure 4.3 and Figure 4.4. This fast shut-off of calcium influx is needed if the model is to reproduce the tight temporal sensitivity thought to occur in STDP (Bi and Poo, 1998). The original plasticity model that this model is based upon (Shouval et al., 2002) accomplishes a similar feat by introducing a back-propagating action potential which transiently raises the membrane potential to increase calcium influx for a short period.

4.4.2 The plasticity model

The model of synaptic plasticity used was based upon the calcium control hypothesis (Shouval et al., 2002), the idea that changes in synaptic plasticity are largely caused by changes in calcium concentration, with high levels of calcium resulting in long term potentiation of the synapse and intermediate levels of calcium resulting in synaptic depression (see Figure 4.6).

$$\Omega = 0.25 + A * B \tag{4.18}$$

$$A = \exp(\beta_2 * ([Ca^{2+}] - \alpha_2))/(1 + \exp(\beta_2 * ([Ca^{2+}] - \alpha_2))) - 0.25 \tag{4.19}$$

$$B = \exp(\beta_1 * ([Ca^{2+}] - \alpha_1))/(1 + \exp(\beta_1 * ([Ca^{2+}] - \alpha_1))) \tag{4.20}$$

$$\eta = 1/(P_1/(P_2 + [Ca^{2+}]^{P_3}) + P_4) \tag{4.21}$$

$$dW_j = \eta([Ca^{2+}]_j) * (\Omega([Ca^{2+}]_j) - W_j) \tag{4.22}$$

The equations above show how the change in synaptic weight W_j is calculated from the calcium concentration. Ω represents the relationship between spine calcium concentration and the direction of plasticity. For low levels of calcium concentration there is no change, for moderate levels of calcium we see depression, and for high levels of calcium there is potentiation. This relationship, given by Ω, is shown in Figure 4.6. The term η represents an additional calcium dependence of the plasticity — at low levels of calcium plasticity changes are weak, but as calcium levels rise, plasticity changes occur more rapidly (see Figure 4.7). The change in synaptic weight dW_j is then given by the multiplication of these two calcium-dependent functions.

This model was chosen as the plasticity model because it was simple to implement and widely used in the literature. Since the model was published there has been some debate about the late LTD which appears to be predicted by this model, but has not been observed experimentally (see Figure 4.12). It was decided to use this plasticity

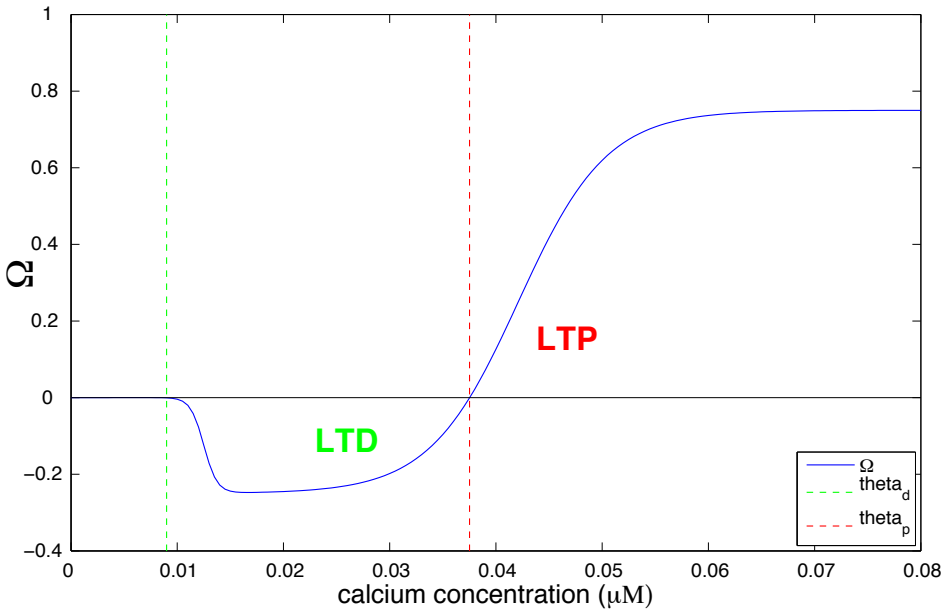

Figure 4.6: The relationship between Calcium concentration and change in synaptic weight (Ω). At low concentrations of Calcium there is no change, at intermediate levels depression, and at high concentrations there is potentiation. The green and red lines show θ_d and θ_p respectively. θ_d is the calcium concentration below which there will be no synaptic change, and above which there will be depression. Once calcium concentration reaches θ_p we will begin to see potentiation.

model despite the problem as it was felt that the presence of late LTD would not effect the questions we were asking about the relative effects of dopamine modulation. It was believed that the magnitude and timing of the late LTD would be parameter sensitive, and therefore might not always be observed in experiment. This appears to be supported by the results in Figure 4.16 which show that the small parameter changes brought about by dopamine modulation can mean that late LTD is not observed.

Other plasticity models were considered, such as (Rubin et al., 2005), but the model by Shouval et al. (2002) was chosen because of it is conceptually clear - a factor which would simplify the analysis and implementation.

4.4.3 Simulated dopamine modulation

In order to quantify the effects of dopamine on synaptic plasticity two major contributions need to be taken into account

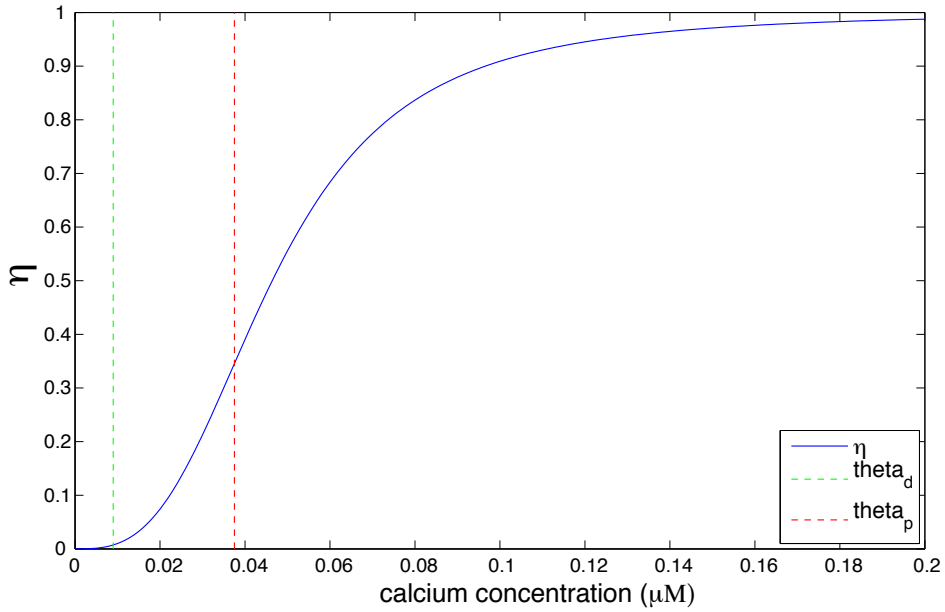

Figure 4.7: η controls the rate of change of synaptic weight. The effect of η is such that there is very little synaptic weight change at low levels of calcium, and a fast rate of weight change at high calcium concentrations.

1. the effect of dopamine on intrinsic excitability of the neuron

2. the effect of dopamine upon synaptic conductances.

To quantify both of these aspects of dopamine modulation I used data from the PhD thesis of Ullrich Bartsch, working in the laboratory of Daniel Durstewitz.

The data consisted of the firing frequency of 30 neurons when injected with a range of step currents in the presence of synaptic blockers. Graphs of nine of these neurons are shown for illustration purposes in Figure 4.8. The step protocol was carried out under three conditions — control, a bath solution of 50μm SKF38393 (a D1 receptor agonist), and a bath solution of 10μm Quinpirole (a D2 receptor agonist). The three conditions were intended to isolate the effects of D1 and D2 receptor stimulation against a background of zero dopamine modulation. The firing rate of each of the neurons was measured using an experimental protocol whereby neurons were injected with a step current for 25 seconds. The initial 20 seconds were ignored to remove transients, and the firing rate was calculated from the remaining 5 seconds. This was repeated 10 times at increasing levels of current injection. Between each condition there was a 10 minute pause to wash out the old agonist and wash in the new. The

conditions were done in the order of control, D2 agonist, D1 agonist, as the D1 agonist was found to cause some long-lasting changes that were not reversed by D2 activation.

A sample of the results from these experiments can be seen in Figure 4.8. The f-I curves of the neurons showed considerable variance, but the results were similar to those obtained by another study of Thurley et al. (2008), which used a slightly different protocol.

The parameters for computationally simulating D1, D2, and control situations were fitted by simulating the same protocol as had been done experimentally, and using an automated parameter search to produce model f-I curves that closely fit the experimental f-I curves shown in Figure 4.9. The parameters that were modified to simulate dopamine modulation are shown in Table 4.1. It is important to note that this model cannot be said to be a model of dopamine modulation itself, as the action of dopamine is slightly more complex than merely being the sum of the effects of D1 and D2 agonists.

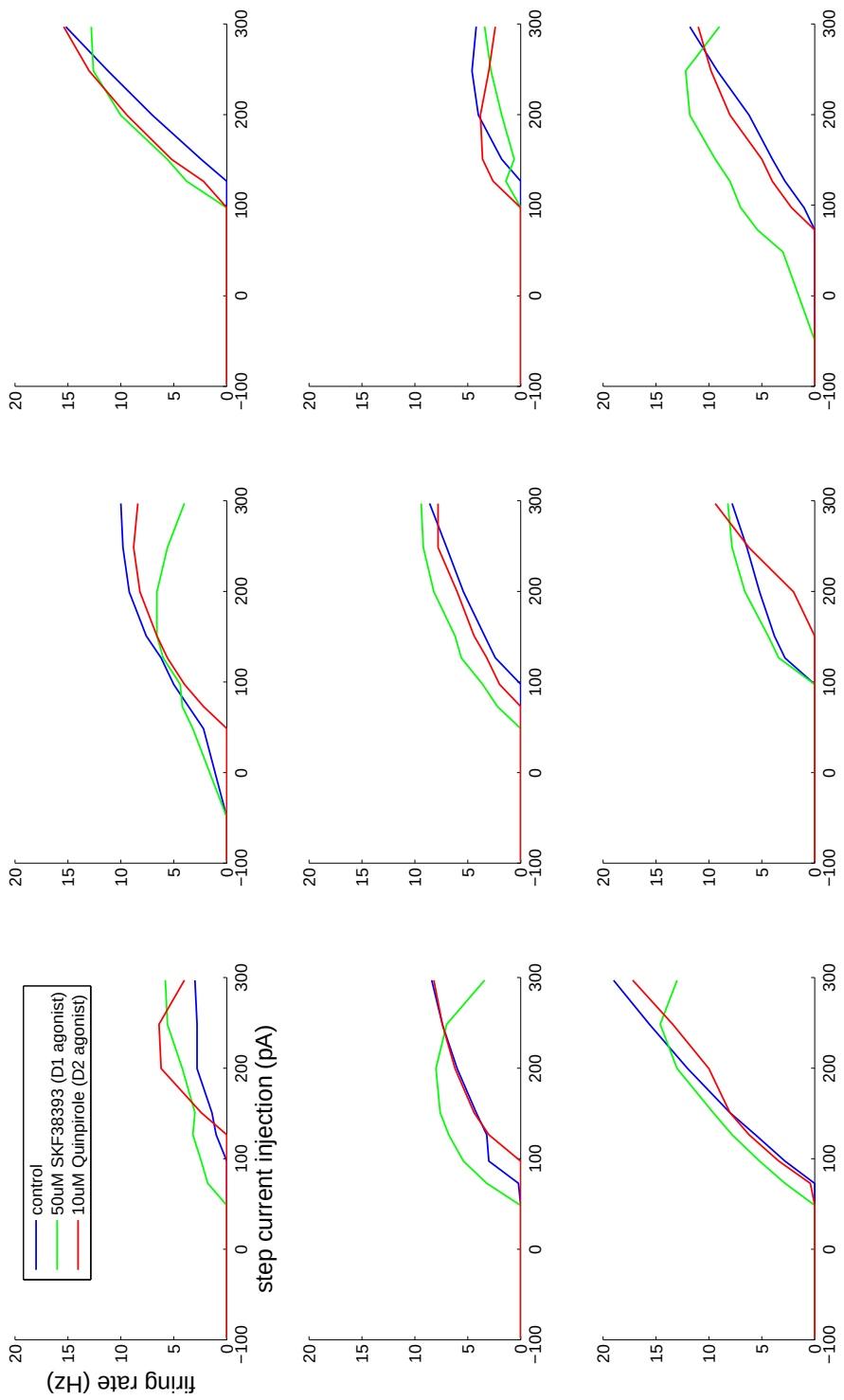

Figure 4.8: The original data from in vitro recordings of prefrontal cortex neurons under modulation by different agonists — note the heterogeneity in neural response. (Bartsch, PhD results).

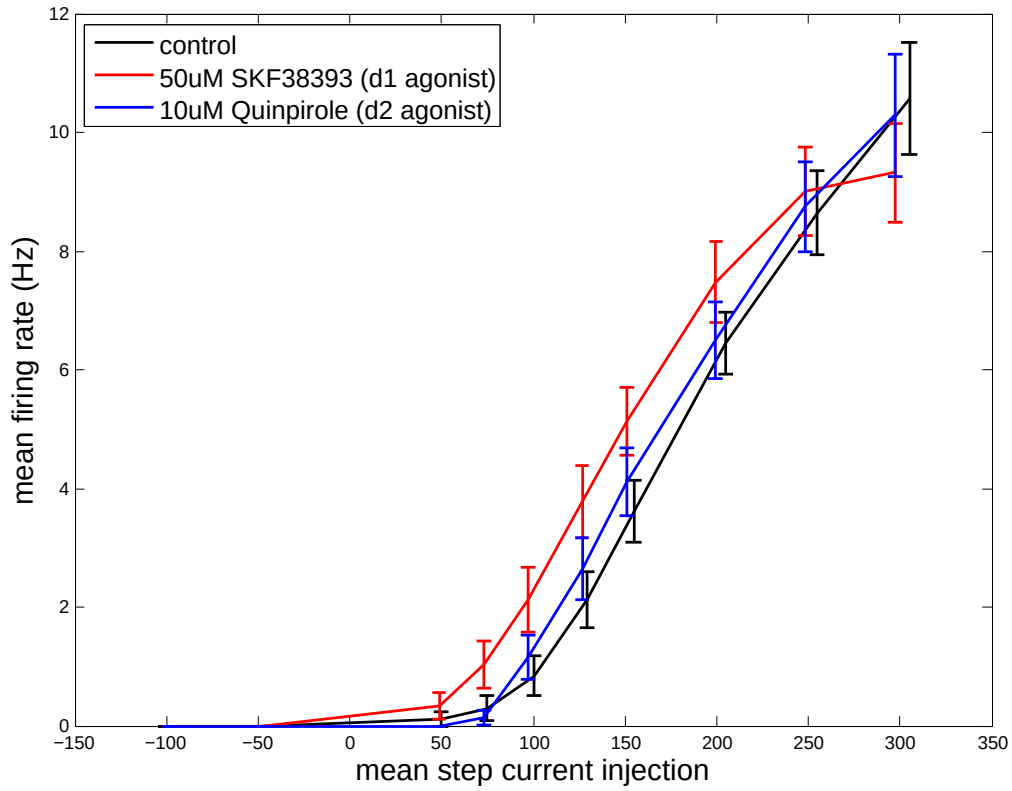

Figure 4.9: The experimental data — the effect of D1 and D2 agonists on firing frequency for a given step current injection (nA). The results shown here represent the mean frequency response across all neurons tested. The significance of the data was calculated using a linear regression. D1 shows a significant difference from control with p=0.05, whilst D2 is significant only at p>0.1. The results indicate that a D1 agonist tends to make the neuron more excitable and lowers the rheobase.

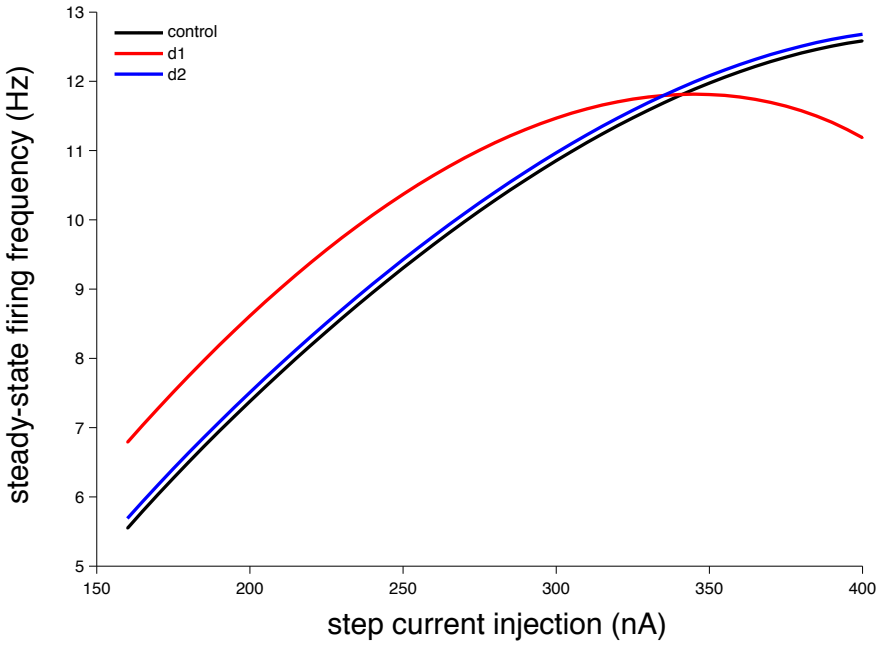

Figure 4.10: The simulated f-I curves. The parameters of the model neurons were modified so as to provide the best possible fit with the in vitro data.

It was not possible to produce model f-I curves that fit the experimental f-I curves over the whole range. For most parameter values, curves that fit at high current input resulted in neurons which failed to show any spiking at low current injections. This is a consequence of opting to use the simple model neuron described in section 4.4.1. This neuron is an accurate model of the membrane trace, but it was later found to be a poor model of the f-I curve at low current injections. This is because the reduced model relies upon a bifurcation to move from silent to spiking, and this prevents it from demonstrating a gradual rise in spike rate as current is injected. Unfortunately there is no choice of parameters that can change the qualitative dynamics of the neuron without losing its ability to fit the membrane trace. The other option was to return to the drawing board with the neuron model to implement more terms to change the dynamics. It was decided it was not a worthwhile time investment to develop a new neuron model to produce a better fit of one portion of the f-I curve.

This issue with the f-I curve at low current injection while significant, was not large enough to effect the overall result. Given other, more significant weaknesses in the model it was decided that this problem was better addressed in future work.

The difference between the experimentally measured f-I curves and the model f-I curves obtained by parameter fitting can be seen in Figure 4.11. Below a current

	$gNMDA_{max}$	$gAMPA_{max}$	gM	gNa
control	0.1429	0.1	19	23.3
D1 bath	0.2	0.1	10.6	21.9
D2 bath	0.0857	0.0857	18	23.3

Table 4.1: The synaptic and neural parameters changed by dopamine modulation, here shown with their specific values under each condition.

injection of 150nA, the model neurons suddenly stop spiking. This is a consequence of the bifurcation needed to generate the spikes.

In addition to the effect of dopamine modulation on neural excitability, we also characterised the effect of dopamine on synaptic conductances. Based upon data obtained in previous studies (Seamans and Durstewitz, 2008), we assumed that D1 modulation increased the conductance of NMDA channels by a factor of 1.4. A D2 agonist decreased AMPA and NMDA conductances by a factor of 0.85, and 0.6 respectively. The details of how these synaptic and excitability changes affect parameter values can be seen in Table 4.1.

The model f-I curves used in this study represent the dopamine modulation used in the original experiment (D1 : 50μm SKF38393, D2 : 10μm Quinpirole), and so this model should only be considered a model of an in vitro situation, rather than in vivo, where dopamine concentrations continuously fluctuate.

4.4.4 The plasticity protocols

In order to look at the full spectrum of effects of dopamine modulation, a selection of plasticity protocols were simulated that would allow for the evaluation of the affect of *frequency*, *pairing*, and relative *spike timing* on synaptic efficacy. The same selection of protocols were used by a previous study, so this made it possible to compare the results of this study with those reported by Shouval et al. (2002). The graphs for this study can be seen in Figure 4.12.

During the course of my preliminary work it was found that the results of this plasticity model were parameter sensitive, and that to reproduce the same results with our model a specific set of parameters was required. However, as the results produced with this model (such as STDP) were felt to be supported by the literature it was decided that this level of parameterisation was acceptable, and so parameters were

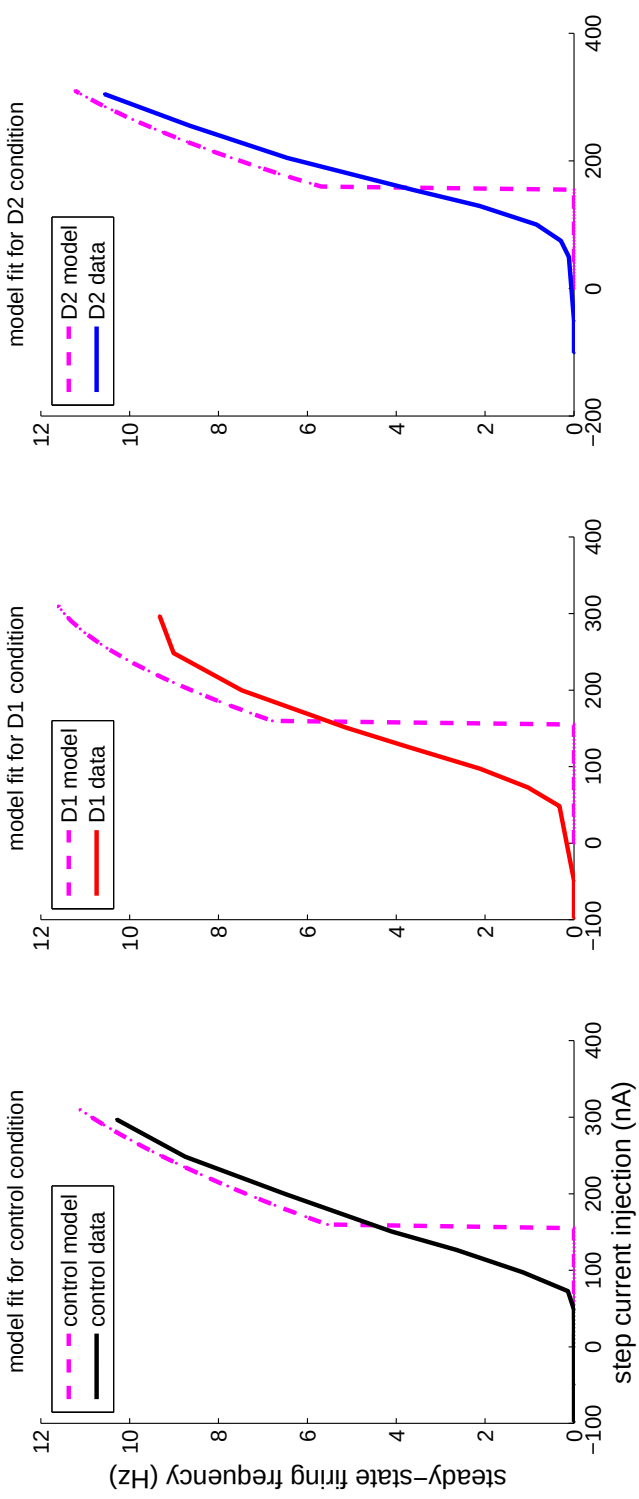

Figure 4.11: A comparison between the model and data f-I curves. The curves produced by the model represent a compromise between sensitivity to small current injections, and a moderate response to higher current injections. If the model neuron is sensitive enough to start spiking at low current injections it shows an overly high firing rate at higher current injections. These kind of compromises are a consequence of designing a simple non-linear model that can fit both the f-I curve and the spike shape — small changes in one parameter can distort the dynamics in another dimension.

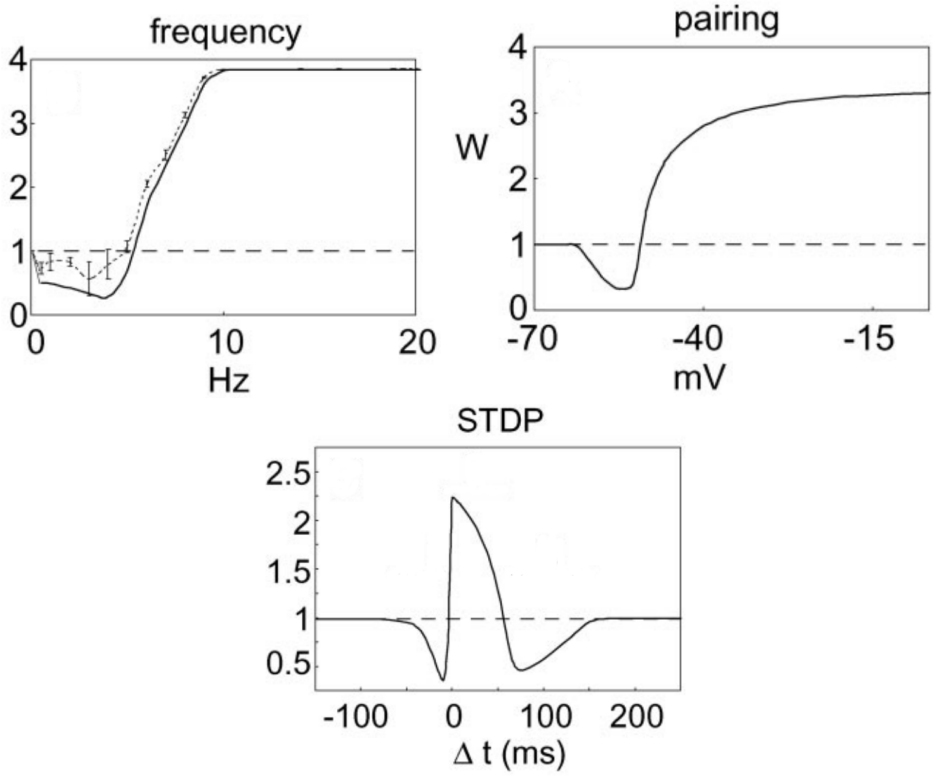

Figure 4.12: Graphs of plasticity curves reported by Shouval et al. (2002). The simulated plasticity protocols used with this model are similar to the ones used in my simulations, so any difference in the results can be attributed to difference in the neuron models and calcium influx mechanisms.

chosen so as to reproduce the basic STDP results in the control case.

The simulated plasticity protocols are designed to mimic in vitro plasticity protocols so that the results of the simulations might be comparable with experimental results. Schematics of the three protocols are shown in figure 4.13

1. The *frequency-dependent protocol* used in vitro involves stimulating a bundle of presynaptic inputs and assuming that it is a population of this bundle that drives the postsynaptic neuron. It is not known exactly how many presynaptic stimuli the postsynaptic neuron receives when this protocol is used, and so in my implementation of this protocol it is estimated that the stimulus recruits nine presynaptic neurons to spike simultaneously. This value was chosen because it produced a response similar to what is observed in experiments. The simulated presynaptic inputs have rise and decay dynamics of AMPA and NMDA channels and are delivered at a systematically varying frequency so as to determine

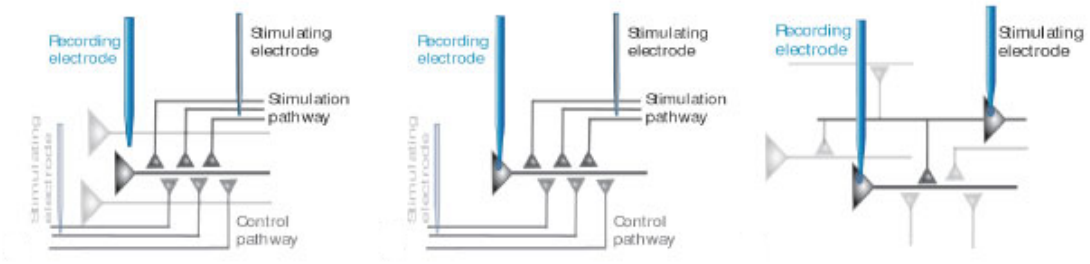

Figure 4.13: Schematics of the three plasticity protocols. On the left is the frequency-dependent protocol, in the centre the pairing protocol, and on the right the timing-dependent plasticity protocol. Adapted from (Shouval, 2007)

the relationship between frequency of input spikes and the eventual plasticity change.

2. Simulating the *pairing protocol* is relatively simple as the mechanisms underlying the method are better understood. In the model the membrane potential of the postsynaptic neuron is clamped at a certain value while the neuron receives presynaptic stimulation. The membrane potential is then systematically varied to determine the effect it has on plasticity.

3. The simulation of the *spike timing-dependent protocol* is relatively simple to relate to the experimental protocol — in the experiment both presynaptic and postsynaptic neurons are stimulated with a current injection multiple times at a chosen frequency. This process is copied in the simulation, and the relative timing of the pre and postsynaptic stimuli are systematically varied to determine the relationship between the timing difference and the eventual change in synaptic efficacy.

4.5 Results

The three protocols were run and produced the results shown in Figures 4.14, 4.15, and 4.16 under simulated D1 and D2 modulation and control conditions.

The results produced by the model are broadly similar to those reported by Shouval et al. (2002) (see the upper graphs in Figures 4.14, 4.15, and 4.16).

The model produces results which suggest that *dopamine modulation of synaptic conductances has a very strong effect on synaptic plasticity*. Dopamine modulation

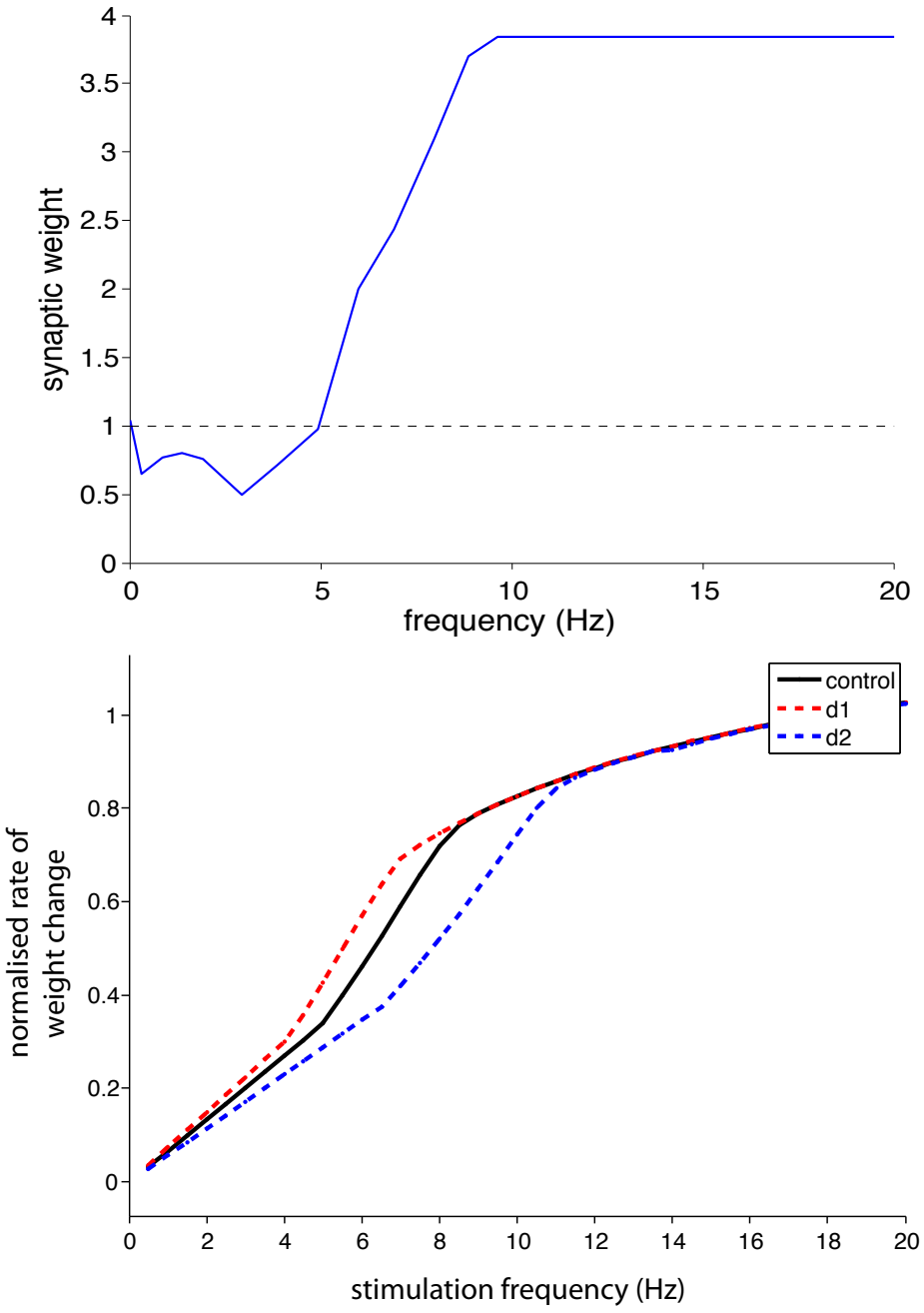

Figure 4.14: A comparison of the results obtained in the frequency protocol with my model (lower), and a previous model (Shouval et al., 2002) (upper). In my model at a low frequency of stimulation there is a small amount of potentiation, and this increases with the presysnaptic stimulation frequency. There is a range within which dopamine modulation can increase or decrease the rate of weight change. Note that the y-axis here denotes that *rate* of weight change rather than the actual weight change. This is because at low frequencies the protocol takes very long, and there is therefore a much larger time in which the plasticity can be consolidated. To make this an even playing field I have instead plotted the rate of weight change.

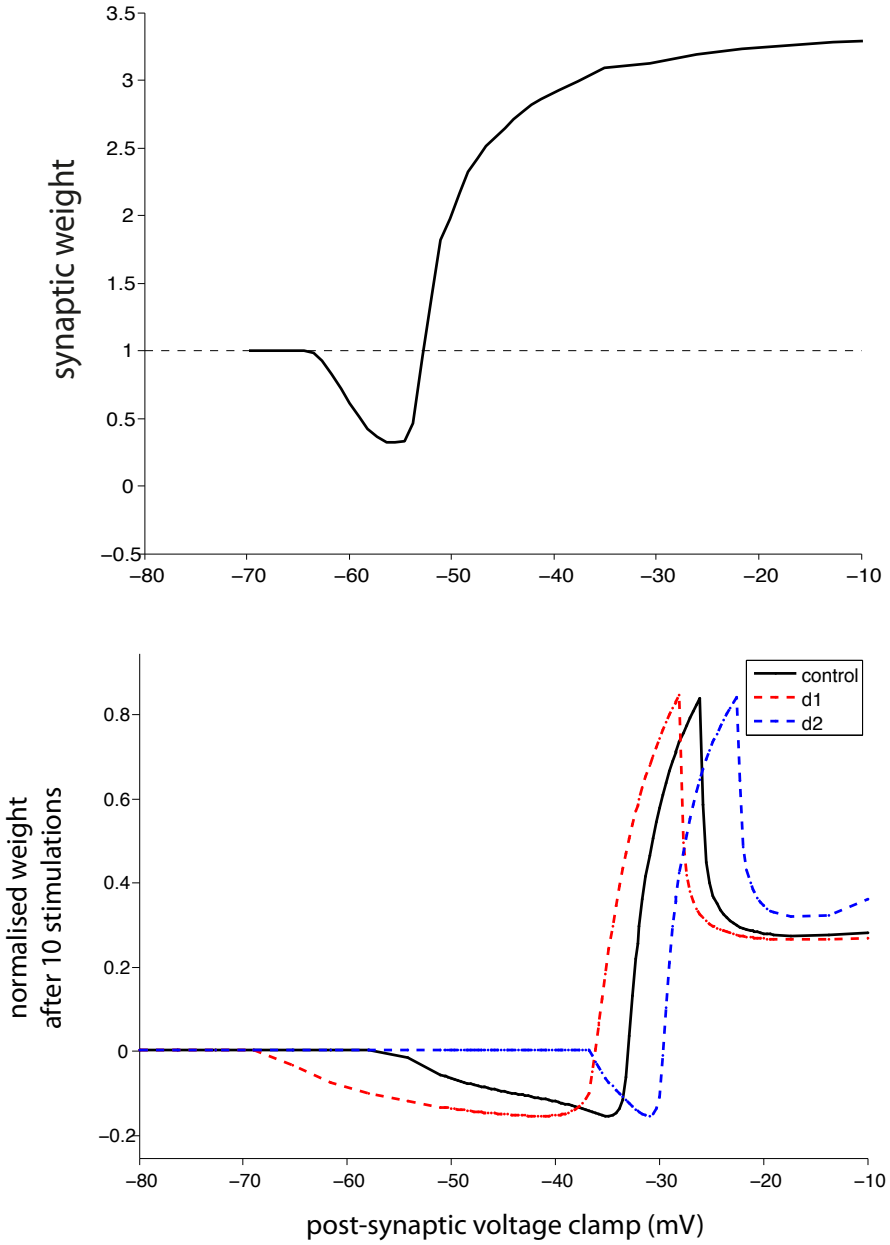

Figure 4.15: A comparison of the results obtained in the pairing protocol with my model (lower), and a previous model (Shouval et al., 2002) (upper). In my model at low post-synaptic voltages initially there is a depression, but as the voltage crossed a threshold, the level of calcium in the spine enters into the LTP region of the Ω curve, and the presynaptic stimulation results in potentiation. Again, D1 and D2 modulation serve to shift the calcium concentration up or down, and this results in the neuron crossing into the LTP region at a lower and higher voltage respectively.

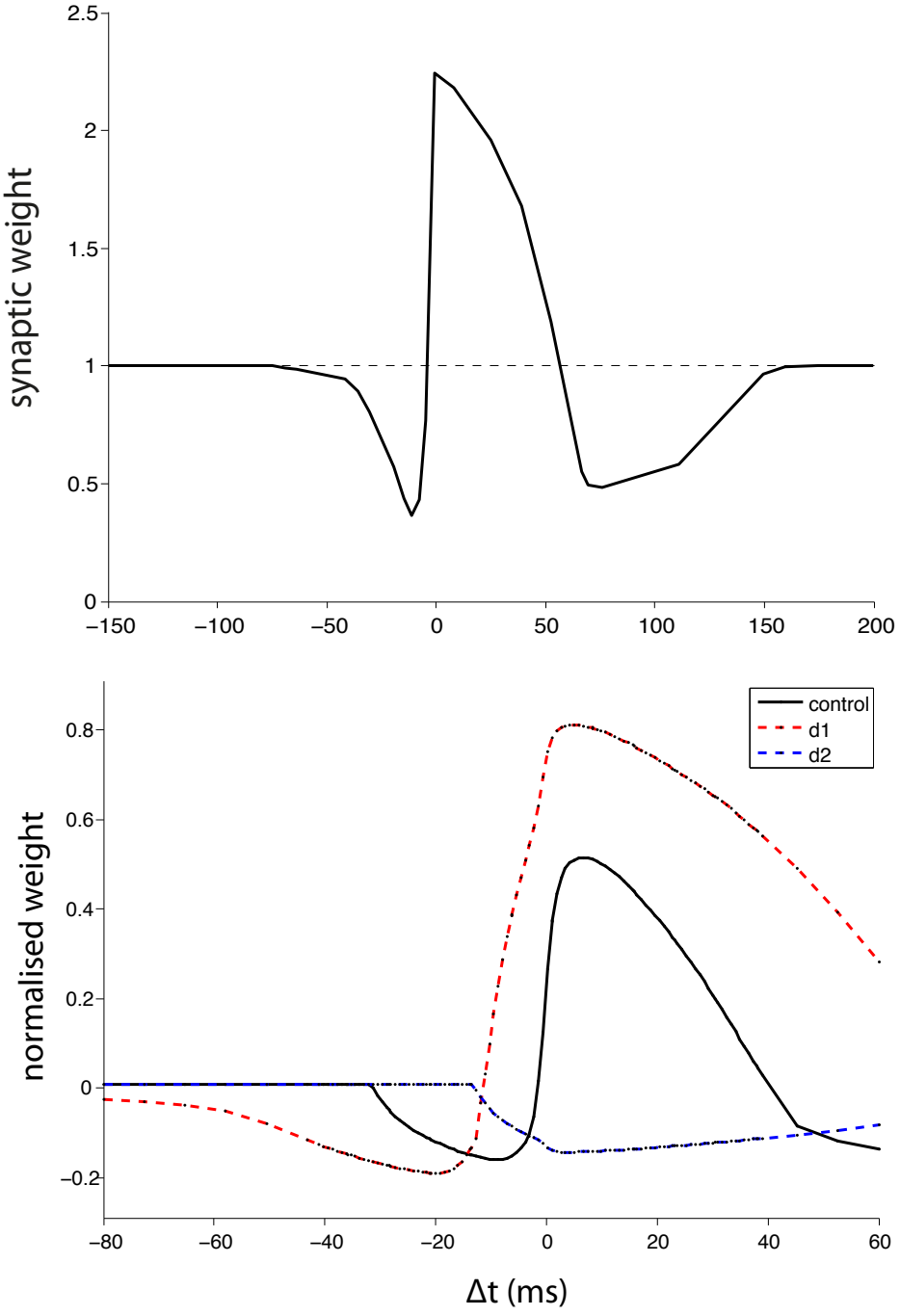

Figure 4.16: A comparison of the results obtained in the STDP protocol with my model (lower), and a previous model (Shouval et al., 2002) (upper). My model reproduces the "vanilla STDP" curve in the control case. The parameters (particularly the resting calcium concentration) had to be set to achieve this. Interestingly the modulation of D1 and D2 agonists has unexpected results on the STDP curve. D1 modulation results in more calcium influx, and so makes the neuron sensitive to pre and post synaptic co-incidences over a wider temporal range AND results in an increased plasticity change. D2 modulation on the other hand allows less calcium into the spine, and so sees only LTD because even for pre before post spikes the calcium concentration does not push Ω into the LTP region.

via D1 receptors results in an increased likelihood of potentiation in almost all of the protocols, whilst modulation via D2 receptors decreases the magnitude of potentiation.

According to the model, *the primary mechanism of dopamine's effect of synaptic change is due to its effect on synaptic conductances, rather than the changes in intrinsic excitability* (Figure 4.17).

This is surprising because the effect of dopamine modulation of excitability is much better understood than its direct effect on synapses, which is less well characterised. This result implies that if we wish to understand the role of dopamine in learning (at least via synaptic plasticity), *we should focus more on the direct effects that dopamine has on synapses*, rather than indirect effects on neuron excitability. This is potentially a valuable result as it suggests that we ought to focus experimental work on a different aspect of dopamine modulation than is usually investigated.

If we compare the results with the closest existing experimental data, namely the more recent study by Xu and Yao (2010), we find that the results are qualitatively similar. This experimental study in the prefrontal cortex found that D1 receptor activation facilitated timing-dependent LTP through a synaptic mechanism (cAMP and PKA) in the presence of picrotoxin. Whilst combined D1 and D2 receptor activation broadens the LTP window under intact inhibition. In addition D2 agonists are capable of inducing timing-dependent LTP through their actions on GABAergic circuits, a mechanism which was not included in this model.

4.6 Discussion

The results produced by the model suggest that changes at the synapse are likely to have a more significant effect on plasticity than changes to the excitability of neurons. However, this result is partially dependent upon the plasticity model we have used here, so it is worth asking - "What key components of the model are responsible for the result?"

4.6.1 The neuron model

The results reported - that synaptic effects are more significant than changes to neural excitability - is a reflection of the relative contribution of the two effects that dopamine has on calcium influx. Dopamine modulation of NMDA receptors increases NMDA conductance by a factor of 1.4, whilst modulation of excitability only changes the

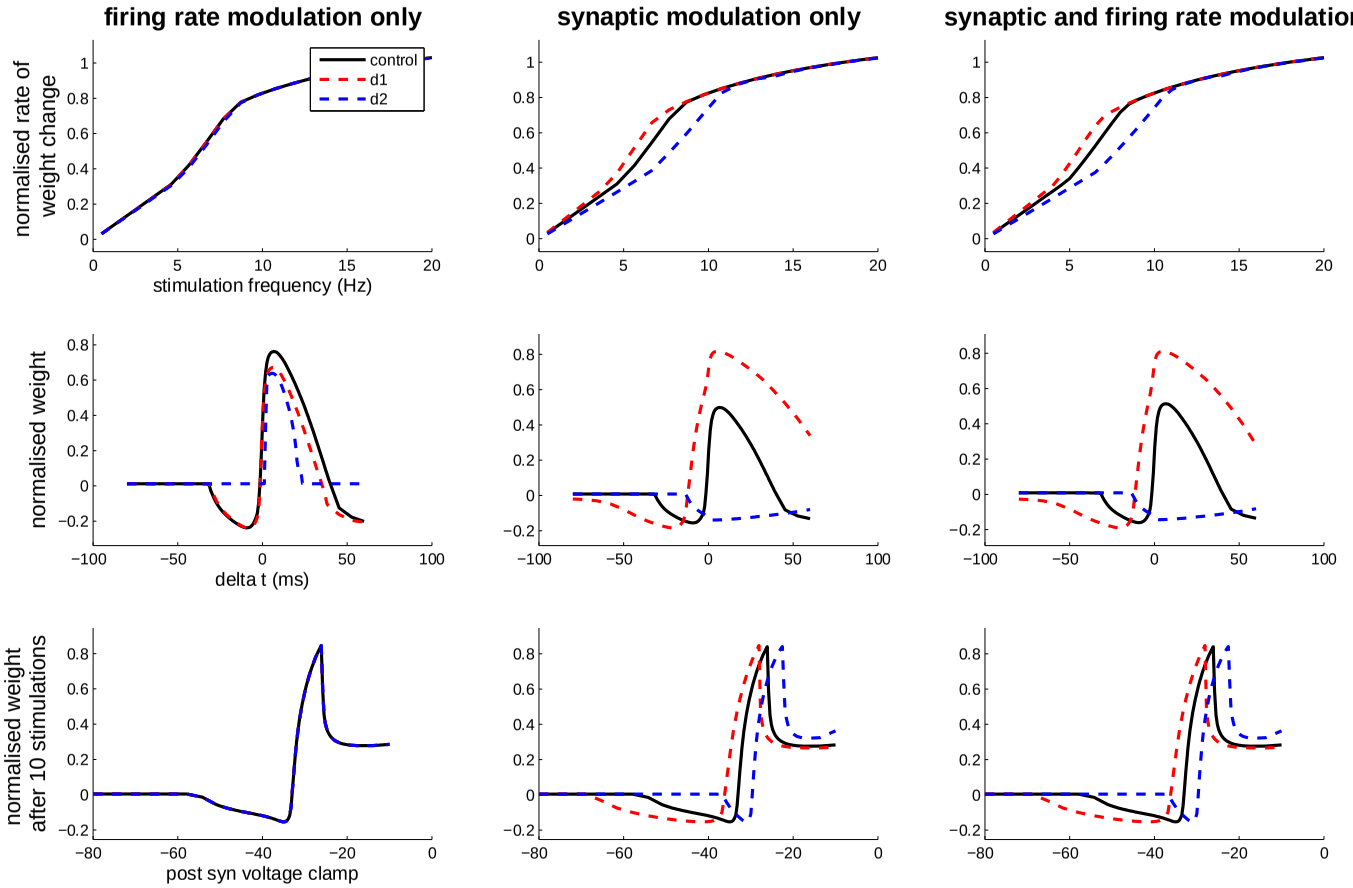

Figure 4.17: Separating out the relative contribution of synaptic modulations and cell excitability modulation to the plasticity results. Simulations were run with modulation of excitability only (changing g_M and g_{Na}, column 1), synaptic modulation only (changing $gNMDA_{max}$ and $gAMPA_{max}$, column2), or both. *The results indicate the the modulation of synaptic conductances is the primary mechanism of the changes in plasticity that are observed* in this model of dopamine modulation of plasticity.

rate at which the neuron spikes. In this model the primary mechanism for calcium influx is the NMDA receptor, and so changes to the conductance of this receptor are more effective at driving calcium input than is a slight increase in the frequency of postsynaptic spikes. *As long as the plasticity rule is calcium based, and calcium influx occurs in this way, then we can expect this observation to hold.*

However, there are two major problems with this scenario. One is that it is known that there are other sources of calcium influx into the synapse other than the NMDA receptor. In addition to calcium influx through the NMDA receptor, neurons also have internal mechanisms which allow for release of calcium from intracellular stores. If we are to verify or extend upon this result we should focus upon accurately characterising these alternate sources of calcium. There are more detailed computational models of calcium influx in the literature, and the findings of these models could be incoporated into future work (Franks, 2001).

Another issue with this model is that the neuron model we have used is a model of the membrane potential dynamics, and this does not necessarily translate into the calcium dynamics of the neuron. It it not clear how realistic it is to attempt to reconstruct intracellular calcium concentration from the membrane potential alone. In addition to the observation that calcium dynamics and membrane dynamics may diverge, it is also possible that dopamine modulates calcium influx independently of changes to the membrane potential. If this was the case we would need new experiments to characterise this effect.

4.6.2 The plasticity model

Aside from issues with the neuron model, the process of implementing the plasticity model made some of its weaknesses clear. Since the development of the plasticity model used in this study it has become more clear that the types of plasticity observed in pyramidal cells is dependent upon the location in the dendritic tree where the observations were made (Kampa et al., 2007).

These results indicate that the dendritic morphology of the neuron has a significant effect upon the outcome of a particular stimulation. Future work could attempt to model the morphology of the neuron, as reconstructions of prefrontal cortex pyramidal cells do exist. However one should be wary of attempting to add an ever-increasing amount of biological realism to a computational model.

In the process of developing the plasticity model it became clear that in order for

the STDP results to be produced, the calcium concentration must be at a fixed concentration when the neuron is at rest - this is highlighted as point A in Figure 4.18. Given the variation that we find in measurements of neural parameters, it is highly unlikely that all neurons maintain this constant value of calcium concentration unless there is a homeostatic mechanism that ensures this is the case.

Another issue that became clear when working with this plasticity model is that there is a problem with the sensitivity to calcium. If the model is to reproduce the well known spike-timing dependent plasticity curve, it must at rest be poised at point A in Figure 4.18, the calcium concentration at which the neuron is exquisitely sensitive to the difference in calcium influx that occurs due to relative timing differences of one pre and postsynaptic pair. However, when a neuron with this kind of sensitivity is subjected to repeated pre and postsynaptic firing, the total level of calcium influx dwarfs the relative difference that occurs during a spike-timing protocol. So, if the neuron is to reproduce the spike-timing results it must be sensitive to changes in calcium of the order of 0.005μM, even though repeated spiking can drive the calcium concentration to increase by 0.1μM. This kind of sensitivity over two different ranges suggests that other mechanisms must be involved.

4.6.3 The methodology

The final issue that I faced when working with this model was to do with the methodology I set out to test at the beginning of this chapter - that of empirical modelling. The level of complexity present in dopamine modulation of plasticity made it difficult to construct a model which could account for all empirical observations. In practice it is not possible to construct a model where all the parameters are based upon observables AND have the model remain ecologically valid. Both the neuron model and the dopamine modulation model had their roots in the data, but the process of combining the two models meant that new variables and mechanisms were required. As the models had been configured to work in different regimes these parameter values had to be set in order to get both models to work together. Unfortunately putting two computational models together does not yield a larger, yet still accurate model. When working with two models with uncertain data the errors propagate and the assumptions multiply. *While there are plenty of ways in which the model could be more technically accurate, the process of developing it has led me to the viewpoint that there are problems with this style of modelling that cannot be completely resolved by adding more*

detail.

4.7 Conclusions

A model of the effect of dopamine on synaptic plasticity was constructed. The model produced results which suggest that dopamine is a potent enhancer of synaptic plasticity, particularly through the effect of dopamine upon synaptic conductances.

However, in the process I became aware of the cumulative affect of the many estimations and assumptions that were required to complete the model. This process of implementing theoretically-derived models, and parameter-fitting was present in every step of the process, despite my best attempts to base the model on firm experimental data.

I have discussed the methodology used in this chapter and concluded that like the engineering-style approach used in the last chapter, empirically-based modelling has it's weaknesses. In the next chapter I will attempt a third approach, which is described below.

4.8 A third approach

My first attempt at answering the question posed at the beginning of this thesis failed because the theoretical models used by what I referred to as an engineering approach were too powerful and could be made to fit any data. Without a firm grounding in experimental data it was very difficult to translate them into experiments that can support or refute them.

My second attempt — the empirical approach described here in this chapter — also ran into problems because I found that it is not possible or desirable to create purely empirical models. Even the experimental observations used to generate the models are theory laden, and these subtle assumptions that are made along the way when constructing experiments become explicit when we try to produce a computational model of the experiment.

Both these attempts ran into problems because of the nature of the question I was asking — *what is the way that the brain solves the problem?* - It is a conclusion of this chapter that it is not possible to answer this question. If our observations are theory laden, then we can never adequately verify whether a computational model is a "true model" of what happens in the brain.

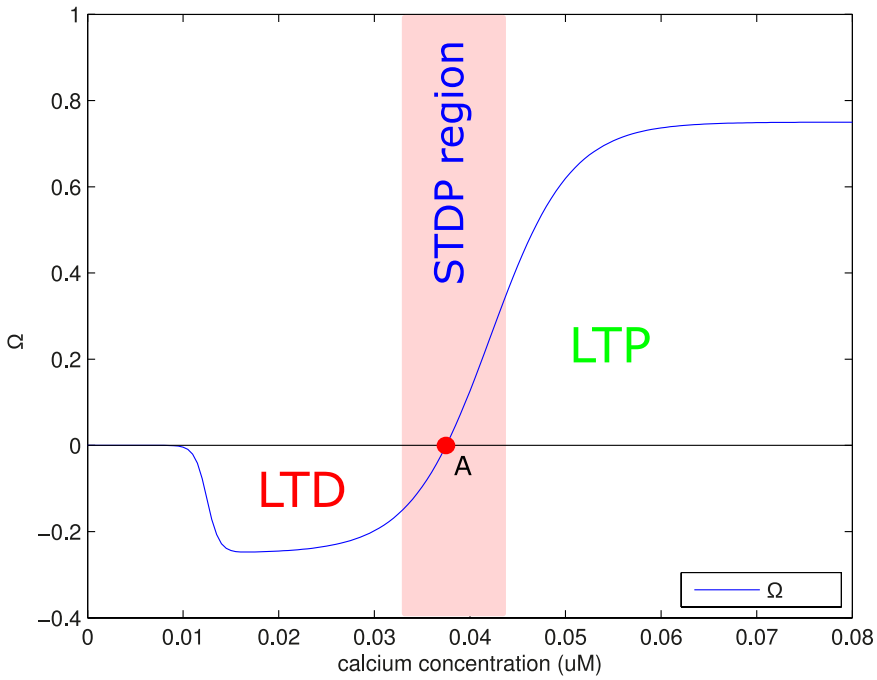

Figure 4.18: For the model to produce depression or potentiation dependent upon the relative timing of pre and post synaptic spikes, the calcium concentration of the neuron must maintain a homeostatic baseline level of calcium at point A. This allows spike pairs with positive Δt to tip the neuron into the LTP region, whilst spike pairs with negative Δt will lead to a net efflux of calcium, and synaptic depression. However, in our model point A is not a stable point, and without an additional calcium homeostasis mechanism to keep the calcium concentration at point A it is highly unlikely the neuron will stay there by chance. Even with the addition of a calcium homeostasis mechanism there is a problem with a plasticity rule based upon this omega function. The STDP region, shaded in the graph above, is very narrow, meaning the synapse must be sensitive to very small changes in calcium concentration. However, this degree of sensitivity is not compatible with the magnitude of calcium that can enter through an NMDA channel if the neuron is repeatedly stimulated.

Rather than launch into a third attempt to overcome this perennial problem, my solution is to avoid asking this problematic question of *"what is really going on in the brain?"*. Instead my third approach will be to focus on providing *an interpretation* or theoretical model of the data that may make some interesting predictions, but does not claim to represent how the brain solves the problem. This method is not interested in how the brain solves the problem, but in whether or not this conceptual model can produce predictions that might prove to be useful. *It marks a change from asking questions about how the brain works, to asking questions about how we can make useful tools.*

Is this kind of modelling useful? Does it have a place within neuroscience? And is it qualitatively different from the two approaches we have already examined, or do they all lie on a continuum?

Chapter 5

Diffusion based Reinforcement Learning model

5.1 Novel contribution

In this chapter I construct a systems-level model of the effects of dopamine on activity in the prefrontal cortex. I show how basic assumptions about the anatomy of the cortex and basal ganglia can lead to TD-like behaviour. The model produces the novel observation that working memory dynamics can be seen as an emergent feature of a reinforcement learning system.

5.2 Introduction

In the previous two chapters I have tried to address the central question of this thesis using a computational model. So far I have looked at:

1. Chapter 3 : An engineering approach — I looked at a model which attempted to show how an interesting theoretical mechanism might be applied to solve a neurophysiological problem

2. Chapter 4 : An empirical approach — I attempted to construct a model based entirely upon empirical data, that would produce quantitative predictions that could be tested by further experiments

Both of these methods met with mixed success. The aim in this chapter is to use what has been learned so far to produce a third model which improves upon, or offers an

alternative to the previous two attempts, AND reflects upon the central question of this thesis — "Does dopamine form part of a reinforcement learning circuit in the brain".

In the last chapter I produced a model which was tied to experimental observations, and therefore potentially useful in making new predictions. However, throughout the modelling process I was forced to make approximations and assumptions as a result of the data that was available. One response to these shortcomings, and indeed the most common response is to try to fix the model by adding more detail... "Perhaps if we implemented a plasticity model which was more aware of the neuron morphology then our model may make more realistic predictions?". The level of complexity that is present in neuroscience almost guarantees that we will never be able to construct perfectly accurate models, so we should consider inaccuracies in our models as inevitable, rather than something we can fix by adding more detail.

One issue that was discussed at the end of the last chapter was that these models both fail in the way that the models try to get at *"what is really happening in the brain"*. The first approach fails because the models it produces are *difficult to verify* experimentally, whilst the second approach fails because it it *easy to verify* and thereby easy to demonstrate its inaccuracy at some higher level of detail. If there are problems at either end of the scale of verification, then perhaps the problem is the process of verification itself? If we relax our need to compare the model to every known experimental observation, if we instead produce models that aim to provide a conceptual framework for interpreting just a subset of the observations, then will the problem go away?

Although this third approach may appear to be just a subtle, semantic difference, it frees us up from having to match the biological data perfectly (a problem in approach 2), at the cost of being able to claim that the model represents what is really happening in the brain (an aim of approach 1 and 2).

So the aim of this chapter is to

- *Interpret the physiological data* about how the brain behaves during reinforcement learning tasks
- *Offer a model* of how dopamine might form part of a reinforcement learning circuit in the brain
- *Suggest future avenues of research*
- *Demonstrate a proof of concept* that the mechanisms in the model can solve the

problem

The key thing to note about this approach is that it makes no claims to represent what is really going on in the brain. In fact it considers the very question to be ill-formed. Instead this should be thought of as a throwaway model, which aims to explain a subset of the experimental data, which in our case are the neurophysiological correlates of reinforcement learning tasks.

If the model we construct is successful in solving the task, then it may prove to be a useful conceptual framework for designing future experiments or clinical interventions, but despite its usefulness it cannot be used to support any claims about what is happening in the biology.

5.3 Aims of the model

The aim of the model in theis chapter is to show how simple assumptions about the effect of dopamine modulation and the anatomy of the cortex and basal ganglia can lead to TD-like behaviour.

It is not an attempt to go the other way - to show how theoretical models from machine learning can be mapped onto the anatomy. The difference between these two approaches is the intended audience - the former is more likely to be of interest to cognitive neuroscientists, while the latter is of interest to the machine learning community.

5.4 Literature review

5.4.1 The neurophysiology of reinforcement learning

Work by Schultz (1998) suggests dopamine neurons fire phasically following an unexpected reward. During phasic firing, dopamine is released from axons in the cortex and striatum (Seamans, 2007), (Arbuthnott and Wickens, 2007). The subsequent increase in extracellular dopamine concentrations and leads to biochemical and electrophysiological changes to neurons in these target regions. The prefrontal cortex in particular is thought to be involved in working memory (See (Goldman-Rakic, 1995) for a review), and it is believed that the dopamine modulation plays a key role in this. In recent years computational models of the prefrontal cortex have suggested electrophysiological mechanisms by which dopamine might facilitate the persistent of delay-period activity. Models by Durstewitz et al. (2000a) have suggest that dopamine

modulation leads to a bifurcation in the membrane potential of pyramidal cells, and this bistable state may lend properties which aid working memory circuits (Durstewitz et al., 2000b). Another model by Tegnér et al. (2002) suggests that working memory comes about by recurrent, reverbatory activity in prefrontal circuits. Other models have proposed that working memory comes about through the bistable dynamics of medium spiny neurons in the striatum (Gruber et al., 2006).

Neurons in the prefrontal cortex that are modulated by dopamine project back to the VTA (Gariano and Groves, 1988) and also to neurons in the striatum (Voorn et al., 2004). These cortical projections which synapse in the VTA are capable of triggering burst firing of dopamine neurons when stimulated (Gariano and Groves, 1988). These reciprocal links off one pathway by which the reward prediction error signal from dopamine neurons could result in future reward prediction error signals.

Projections from the cortex to the striatum converge on synapses at medium spiny neurons - gabaergic neurons which in turn project to the substantia nigra and ventral tegmental area (Voorn et al., 2004). Direct projections from these areas are known to inhibit VTA neurons, and prevent phasic bursts (Diana and Tepper, 2002). The striatum is densely innervated by dopamine axons and forms another loop by which a reward prediction error signal can contribute to future estimates of reward prediction error. It has been found that a proportion of these inhibitory projections carry information about the degree of expectation of reward. In a study by Hollerman et al. (1998) 250/1500 neurons sample in the caudate nucleus, putamen, and ventral striatum showed activity that reflected the expectation of experimenter instructions or presentation of a trigger stimulus.

Outside of the striatum, cholinergic neurons in the PPTg project to the VTA and are thought to relay primary sensory input of rewarding stimuli (Grace et al., 2007). These cholinergic neurons are particularly effective at triggering phasic bursts of dopamine neurons.

5.4.2 Existing computational models

The first model to make a connection between models of reinforcement learning and dopamine was that by Houk et al. (1994). They proposed that circumscribed regions of the striatum called strisomes, and neighbouring matrix regions formed the basis of a reinforcement learning circuit in the basal ganglia. Houk et al. (1994) described their model using the “actor-critic” terminology used in machine learning, and suggested

the striatum was the actor, and dopamine neurons the critic. Similarly to the model by Izhikevich (2007) described in Chapter 3, Houk et al. (1994) proposed that dopamine reinforced biochemical eligbility traces such as CAMKII. A later model by Suri and Schultz (1998) implemented an actor-critic model and demonstrated that it could successfully back-propagate reward prediction error. Contreras-Vidal and Schultz (1999) based their model on Adaptive Resonance Theory (ART) and suggested that prefrontal cortex activity guides short and long-term processing in cortico-striatal circuits. Brown et al. (1999) also proposed a biologically-inspired model of reinforcement learning which they claimed offered an alternative to TD formulations. Another model by Berns and Sejnowski (1998) provided a systems-level description of how the basal ganglia might implement action selection. A review of biologically inspired reinforcement learning models can be found in models can be found in (Bar-Gad et al., 2003).

5.4.3 The link between working memory and reinforcement learning

The only other article I have been able to find making a link between reinforcement learning and working memory is by (Savin and Triesch, 2009), who also suggest that the link has so far not been recognised.

5.4.4 The neurophysiology of reinforcement learning : key observations

As stated above, my aim is to interpret the physiological data — the key observations I intend to build my model on are

1. When released dopamine changes the excitability and synaptic conductivity of neurons with dopamine receptors. According to the literature, this effect of dopamine upon cortical pyramidal cells can be separated into two timescales. In the first 100-200ms VTA activity evokes an EPSP and inhibits spontaneous firing (Gorelova et al., 2002), whilst over a longer time period (around 30 minutes) the effect of the dopamine that is released is to increased evoked spiking (Lavin et al., 2005) and to enhance effective AMPA and NMDA synaptic conductances (Seamans et al., 2001).

2. Dopaminergic neurons appear to spike in patterns that correspond to the reward prediction error signal described by the temporal differences algorithm.

(Schultz, 1998).

The central question in this thesis is about the relationship between these two observations. Is phasic dopamine the cause of reinforcement learning? How might observation 1 lead to observation 2? Is the action of dopamine on cortical and striatal neurons enough to lead to the reward-linked patterns of spiking observed in the dopaminergic nuclei?

5.4.5 The role of dopamine in a reinforcement learning circuit

The model I looked at in chapter 3 Izhikevich (2007) attempted to solve this problem by expressing the action of dopamine as a change in plasticity at individual synapses. This made the synapses more sensitive to the neuron spike rate and the relative ordering of pre and post synaptic spikes. Izhikevich suggested that this mechanism was enough to solve the distal reward problem. But, as we have shown in chapter 3, there are some problems with the model.

One of the key issues that was found with this model is that it was sensitive to changes on a very short timescale — the millisecond timescale of individual action potentials. This is in contrast with the timescale of the behaviour that determines whether or not there is a reward (seconds and minutes). According to the separation of timescales argument laid out in section 1.4, if we are looking to find the proximal cause of some effect we should start by looking for phenomena that vary on the same or a slightly shorter timescale, *before* turning our attention to more distal causes. Models which attempt to explain macroscopic phenomena using microscopic mechanisms are jumping the gun, and are likely to run into problems similar to those we observed with the engineering style of modelling in chapter 3 — *if we are free to pick and choose our microscopic mechanisms, then we can probably put them together in many ways to solve the problem, none of which necessarily represent the way the biology solves the problem.*

To avoid running into this issue I intend to build a model from phenomena of a similar temporal and spatial scale as the behaviour we are trying to explain. Thus, rather than explain behavioural learning using millisecond phenomena at individual synapses, I will focus upon the effect of dopamine over seconds and minutes on the mass activity of *populations* of neurons.

5.5 Constructing the model

Based upon the reasoning outlined above, the models in this chapter will be based upon phenomena that occur at a slower timescale than the single neurons and action potentials that are often the basis of models in computational neuroscience. In constructing the model I will make a few assumptions. To make these explicit and clear they are:

1. A phasic burst of dopamine neurons that occurs after an unexpected reward leads to a transient increase in dopamine concentration (Arbuthnott and Wickens, 2007).

2. The short increase in dopamine concentration changes the state of neurons for a period longer than the elevation in concentration (Lavin et al., 2005).

3. A brief phasic increase in dopamine concentration causes neurons to enhance their excitability (Lavin et al., 2005), which over a long period results in enhanced plasticity and spine growth at the synapses of active neurons.

4. These changes lead to a greater degree of recurrent activity among neurons that were active at the time of reward relative to those which were not active

These assumptions have been chosen for the model because they are a) fairly simple, b) supported by the experimental data, c) offer properties that might be useful for solving the distal reward problem. In order to show how these mechanisms might play a part in a reinforcement learning circuit in the brain, I will construct the reinforcement learning model in 3 steps

1. First I will show how the plasticity changes that occur after a reward can translate into persistent delay-period activity (working memory).

2. Secondly I will show how the enhanced activity that occurs at the onset of predictive cues can be learned as an early prediction of reward

3. Finally, I will suggest how the development of an expectation of reward can lead to a back-propagation of reward prediction error and the eventual firing of dopamine neurons in accordance with the predictions of reinforcement learning.

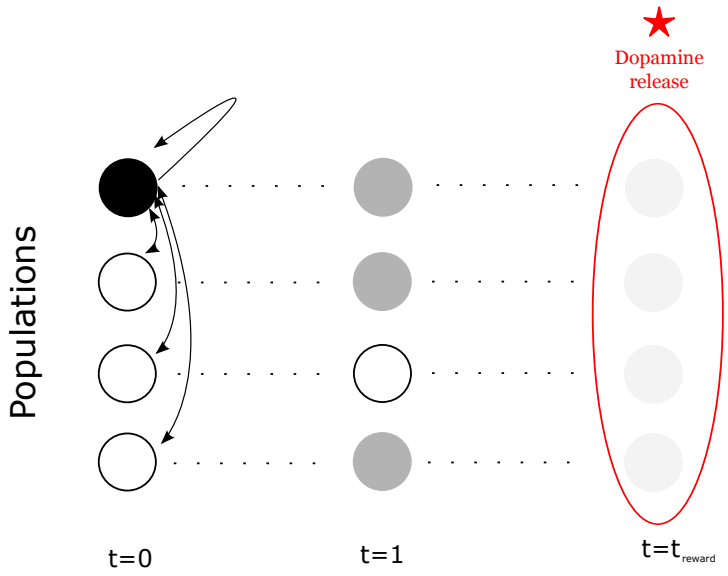

Figure 5.1: A schematic of the model network showing populations of neurons during the course of a single trial. The shading of the populations indicates their activity, so the first population starts as highly active, and during the course of the trial its activity diffuses to the other populations. In the simulations one population is chosen as a cue, and starts with a correspondingly high level of activity. At the end of the trial a reward is presented (simulated by the learning rule) if the cue was presented. In this model a cue is presented in the first timestep (by increasing the activity of the cue population), and during the remainder of the trial the activity is free to diffuse through the network. The activity of the populations at the time of reward is used to update their recurrence rates — this process is supposed to be analogous to the activity-dependent plasticity that occurs when dopamine is released.

5.6 Learning working memory from reward

5.6.1 Method

Behavioural stimuli in reinforcement learning experiments are thought to be represented by populations of neurons rather than single units, and so my models are composed of populations of neurons, rather than single neurons. In this model I am not particularly concerned about the exact millisecond dynamics of individual neurons, and so the activity of the population is represented as a number between 0 (no activity) and 1 (highly active).

Populations are connected to one another by thousands of synapses – the value assigned to these connections represents the *relative number of synaptic connections*.

Following Compte et al. (2000), populations are more likely to be connected to their neighbour than distant populations, such that the distribution of connections between populations can be represented using a Gaussian function.

$$W_{ij} = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-d_{ij}^2}{2\sigma^2} \tag{5.1}$$

Here d is the distance between two populations when the populations are laid out on a ring. For example, in a network with 100 populations, the distance between population 1 and 100 is 1. The parameter σ controls the width of the gaussian distribution.

The number of internal connections *within* a population is variable, and this number is referred to as the recurrence rate (RR). At the end of each timestep in the simulation, a fraction of the activity of each population escapes and is free to diffuse into other populations. The fraction of activity that escapes with each timestep is $1 - RR$, implying that populations with a larger degree of recurrence hold onto their activity for longer. As the escaping activity merely diffuses into another part of the network, the overall activity in the network is conserved. To simulate the process of neural diffusion, for each population, two recipient populations i are chosen according to a random weighted sample, where the weights for the sample given by the interpopulation connection weights W_{ij}. This additional activity passed on from the jth populations to the two recipient populations i is given by

$$AdditionalActivity_i = PopulationActivity_j * \frac{(1-RR)}{2} \tag{5.2}$$

The above rule effectively transfers the activity lost by the population j to the two recipient populations. In this simple model I am interested in what will happen in a simple reinforcement learning scenario, such as the one illustrated in Figure 5.1. In this example a cue is presented at $t = 0$, and paired with a reward at $t = t_{reward}$. In this first model, at the time of reward, a variable representing dopamine concentration is manually increased, and I examine the effect this has on the populations that are often active in the timesteps before the reward.

It was proposed in my assumptions that the effect of a phasic increase in dopamine concentration is to potentiate synapses between neurons that were active at the time of reward. In biology this is a slow process which happens in the seconds and minutes after a reward, but one that will be approximated here by increasing the recurrence rates of the populations that were active at the time of reward using the following formula:

$$RR_{n+1,j} = \left(1 - RR_{decay}\right) * RR_{n,j} + \left(PA(t_{reward})_j - \overline{PA(t_{reward})}\right) * \triangle RR$$

Here $RR_{n+1,j}$ represent the value that the recurrence rate will take on the n+1th trial due to the reward-based learning that takes place at the end of the nth trial. RR_{decay} is the rate at which the recurrence of a population decays over time independently of learning. RR_j represents the recurrence rate of the jth population, and *PA* equals the Population Activity. The effect of the *PA* term is that populations with greater than average activity at the time of reward will see their recurrence rates increase, whilst those lower than average will see their recurrence rates decrease. This serves to implement some basic credit assignment. $\triangle RR$ is the rate at which the recurrence rate changes during learning. On trials where there is no reward, no explicit dopamine release is modelled, and so only the first, decay part of the equation is relevant.

During the simulations, 15 trials were run where the cue and reward are paired, seperated by 15 trials where there was no cue and no reward. In the model a presentation of the cue is simulated by artificially setting the activity of the randomly chosen cue population to 1. The recurrence rate of all the populations starts at a uniform value of 0.05.

5.6.2 Results

The results from this first phase of the model are shown in Figure 5.2. As can be seen in the figure, the repeated pairing of cue and reward causes an increase in the recurrence rate of the cue population. This can be seen more clearly in a plot of how the recurrence rates change with each trial (Figure 5.3).

After 15 rewarded trials we can see that the recurrence rate of the cue population has increased dramatically. Interestingly, the recurrence rates of the populations neighbouring the cue population are also increased by an amount proportional to their distance from the cue population. This property is known as *generalisation*, and in most circumstances is a very useful property, particularly in situations where the reward contingencies of stimuli are continuous in space - ie. a similar stimulus is more informative than a random stimulus. Cortical columns are one example where features are represented spatially through the cortex, and in some parts of the prefrontal cortex, reward contingencies are also represented in a similar manner (Rao et al., 1999). One major advantage of generalisation is that the brain is robust to the loss of individual neurons, or even populations.

Figure 5.2: In early rewarded trials the recurrence rate of the cue population (population 1) is not strong enough to maintain peristent activity. However, as the cue and reward are repeatedly paired, the effect of learning is to increase the recurrence rates of the populations which reliably precede reward. At the end of the simulation the recurrence rate of the cue population is high enough for that population to demonstrate persistent activity that is reminiscent of working memory.

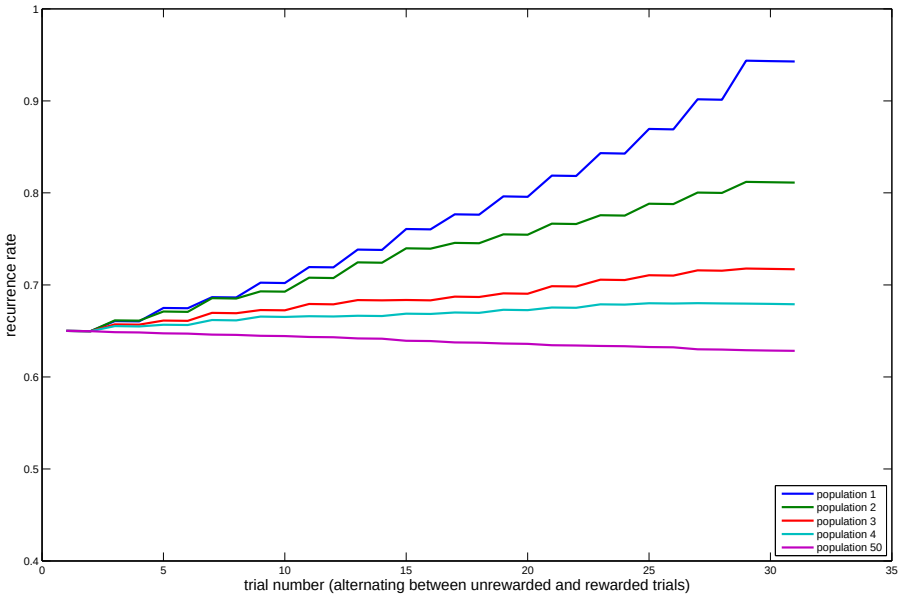

Figure 5.3: The graph shows how the recurrence rates of populations change during the course of the trials. The lines zig-zag because the simulation alternates between rewarded and unrewarded trials. The figure shows how close synaptic connections between population 1 and its neighbours can lead to generalisation — the final recurrence rates are higher for the populations closer to the rewarded cue population (population 1). These neighbourhood relations can lead to generalisation, whereby a similar cue (such as the stimulus represented by poulation 2) can lead to a prediction of reward.

In my simulation the recurrence rates started out with a uniform value, and so the final ranking of recurrence rates according to distance from the cue population shows how reward might cause cortical columns to self-organise so as to represent reward continuums according to distance in the cortex (Rao et al., 1999). [1]

Returning to the results of the simulation, we can see that after 15 cue-reward pairings the recurrence rates have increased considerably. We can see in the last rewarded trial of Figure 5.2 that the model has begun to show some sustained delay period activity in a manner reminiscent of what is observed in working memory tests (Fuster and Alexander, 1971). In fact, in the final state of this simulation, where the model "knows" that when the cue appears that the reward will definitely come — this type of behaviour is identical to working memory. Although it is not often pointed out in the literature, following section 2.4.2 I propose that the behaviour is indistinguishable because *working memory is an emergent property of a reinforcement learning circuit*. Although in this simple model it is only the trace of the cue which is kept active during the delay period, I propose that this same mechanism can explain how goal-dependent motor control can be learned. In a realistic learning scenario, *it is not only the trace of the cue that would be kept active, but the whole sequence of cues and corresponding motor commands that reliably led to reward in previous trials*. It is this sequence that would be learned and played back in precise order during the delay period.

One problem that is visible with these results is that the recurrence rate of the cue population has become so strong by trial 15 that random activity can accumulate during a trial such that the cue population becomes active even when the reward is not presented. This can be seen in the last unrewarded trial of Figure 5.2, bottom right graph, where population 1 begins to accumulate activity as the trial goes on because it has such a strong recurrence rate. This is a problem if we assume that the activity of these cue populations constitutes a prediction of future reward, as such activity would count as a false positive.

[1] Based upon these results we might be tempted to conclude that the brain is optimally evolved for the stimuli, but likewise it could also be the case that the stimuli are classified this way as an by-product of this kind of self-organising wiring process. To ask whether it is the brain connectivity that causes the stimulus statistics, or the stimulus statistic that cause the connectivity is like asking whether the chicken or the egg came first.

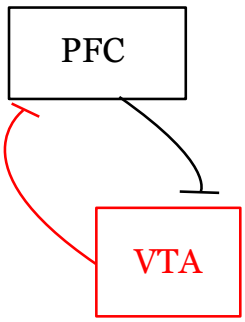

Figure 5.4: The populations in our model represent populations of neurons in the prefrontal cortex. A fraction of the neurons in the populations have excitatory projections to the ventral tegmental area, and so activation of these populations is able to trigger dopamine release, and hence learning. Because the activity of these populations can precede and contribute to a reward, we will assume that the activity of these excitatory projections constitutes a reward prediction.

5.7 Back propagation of predictions

In this section I build upon the model in section 5.6, but this time adding a mechanism to simulate the process of reward prediction more realistically. In the first model I manually increase dopamine release at the time of reward. However, in reality dopamine can be *released before the arrival of the actual reward* if a reward is sufficiently strongly predicted by a cue. But how might this process of prediction occur in the brain, and how can it cause dopamine neurons to fire earlier?

I propose that it occurs due to excitatory connections from cortical regions that act as predictors of reward by stimulating the dopamine neurons, and thus enabling them to fire earlier than the time of reward. It is known that there are profuse projections from the cortex to the ventral tegmental area (Gariano and Groves, 1988), and that the prefrontal cortex in particular represents reward contingent stimuli (Wallis and Miller, 2003). If we assume that the populations in our model are populations of neurons in the prefrontal cortex, and that some of these neurons have excitatory projections to the VTA, then high levels of activity in these populations ought to be able to bring about the release of dopamine in our model. A simple schematic of this architecture, which will be used for the next model, is shown in Figure 5.4. I will propose that **strong activity of these populations constitutes a reward prediction**, and is therefore capable of stimulating a reward prediction error (unexpected reward) and hence dopamine release.

I will incorporate these predictions in the next model by simply assuming that

above a certain threshold, excitatory input from the cortical populations in sufficient to trigger a phasic burst of the VTA.

Ultimately the question of exactly how these excitatory connections affect the VTA must be answered by further investigation of the biology. The firing of dopaminergic neurons is highly complex as it involves an interaction between glutamatergic (AMPA and NMDA), gabaergic, and dopaminergic neurotransmitter, and the natural pacemaker activity of the neuron (Grace et al., 2007). In the model the fixed, per neuron threshold implies that it is primarily strong predictions that cause a phasic burst of the VTA neurons, rather than an additive effect of many weak predictions. To show what happens when reward predictions can trigger dopamine release, another round of simulations was run using this new model. In this simulation, if the activity of the strongest population exceeds the VTA threshold in the timestep before dopamine release occurs (this starts out as the timestep in which the primary reward arrives), then the time of dopamine release is shifted backwards one timestep on the next trial. By this mechanism it is possible for successful predictions to repeatedly backpropagate the firing of dopaminergic neurons. An algorithm for this is given in section 5.7.1.

5.7.1 Method

In this phase the simulation will be the same as in the previous section, only this time rather than the manual dopamine release, the timing of dopamine release will be determined by the following logic.

if $max(PopulationActivity) > 0.4$ AND $t+1 = t_{reward}$

then $t_{reward} = t_{reward} - 1$

Where $max(PopulationActivity)$ will return the activity of the most active population. This will potentially allow dopamine to be release early if there is a strong reward prediction from the cortical populations. The architecture of the network is the same as before, only this time I am explicitly labelling the populations as prefrontal cortex pyramidal cells with excitatory connections to the ventral tegmental area.

5.7.2 Results

The results from this second phase of simulations are shown in Figure 5.5. The graphs indicate that the dopamine that is released at the end of each rewarded trial causes the recurrence rates of the reward-predicting populations to rise, until their activity in the timestep before the reward arrival is enough to trigger a phasic burst of dopamine

Figure 5.5: A similar simulation as shown in Figure 5.1, only this time dopamine release can be triggered if the 'predictions' from the cortical populations are greater than the VTA threshold. The red crosses mark the time at which dopamine release occurs, and so we can see that allowing predictions to trigger dopamine release can result in the cue-reward relationship being learned more quickly. In this example, population 1 shows similar persistent activity in trial 12 as it does in trial 15 in the earlier simulation showed in Figure 5.1.

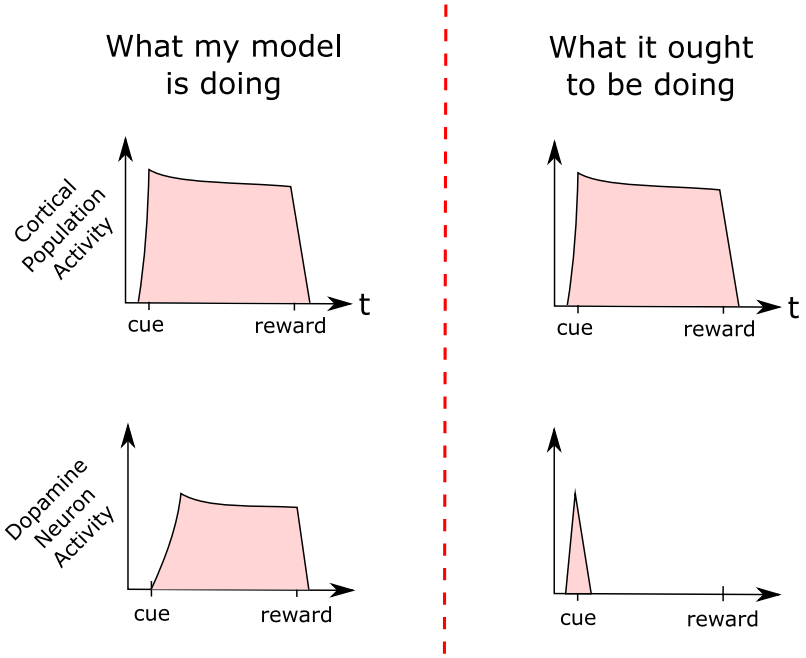

Figure 5.6: On the left the activity of cortical populations and dopamine neurons in the model is shown at the end of the learning process. On the right is what has been observed in animals (Schultz, 1998), (Shuler and Bear, 2006). At present the cortical activity in my model is driving dopamine neuron activity from the moment of the cue presentation all the way until the reward. However, the experiments tell us that the activation of dopamine neurons is transient, and occurs only at the time of cue presentation. This suggests that the model requires an additional mechanism that will silence the dopamine neurons after their initial burst.

neurons before the reward itself arrives. This process is a positive feedback loop — the earlier the reward prediction, the greater the remaining cue trace, and therefore the greater the increase in the recurrence rate. With each repeated trial, the time of reward release is propagated backwards until the onset of the cue alone is enough to trigger the release of dopamine neurons. It is this process which is predicted by reinforcement learning, and this process which is observed in the dopamine neurons of behaving primates.

However, the mechanism described here is still incomplete — at present the "prediction", or prefrontal activity triggered by the cue continues to excite the dopamine neurons *throughout the delay period*. But, from previous experiments we know that the phasic burst of dopamine neurons ought to be transient and only occurs *at the onset of the cue*, and not after it. A schematic explaining this process is shown in Figure 5.6.

In the next section I will address this problem with the third phase of the model. The model has sucessfully solved the problem of backpropagation of predictions, but not backpropagation of *reward prediction error*. In some sense, this can be counted as a successful step in the modelling process — by making one's model of the system explicit, one often finds that the initial model is too simplistic, and another mechanism is required. In the next section a mechanism will be added that will result in the backpropagation of reward prediction errors, rather than just the backpropagation of reward prediction.

5.8 Back propagation of prediction error

The previous section showed how a reward prediction could be backpropagated. However, it is known from experiments that dopamine neurons do not fire phasically after reward prediction, but after an unexpected reward prediction (reward prediction error)

$$RPE = prediction - expectation + actualreward \tag{5.3}$$

In this model I assume that the actual reward signal (information about the sensation of the primary, unconditioned reward) arrives via a subcortical input which has the ability to strongly and quickly activate dopamine neurons when a primary reward occurs. One candidate structure for this is the pendunculopontine tegmentum (PPTg), which has cholinergic projections to the VTA which are capable of causing phasic spikes (Grace et al., 2007).

In the formula above I have suggested that *alongside the reward prediction there is also a reward expectation* which acts to inhibit the dopaminergic neurons, hence the minus sign in equation 5.3. I am proposing this expectation term because it is known that as an animal is repeatedly exposed to the reward it accumulates evidence that the reward will arrive, and after this evidence has been learned, the primary reward has a diminished ability to cause phasic dopamine release via the activity of the PPTg. This learning-dependent inhibition of reward predictions closely mimics our everyday concept of expectation, and so I will refer to it as such. It is this phenomenon that seems likely to hold the key to back propagating ***unexpected*** *reward predictions*, rather than just backpropagating reward predictions.

If this is the mechanism by which reward prediction error is backpropagated, then how might this work in the brain? As I have suggested it is most likely to be the work

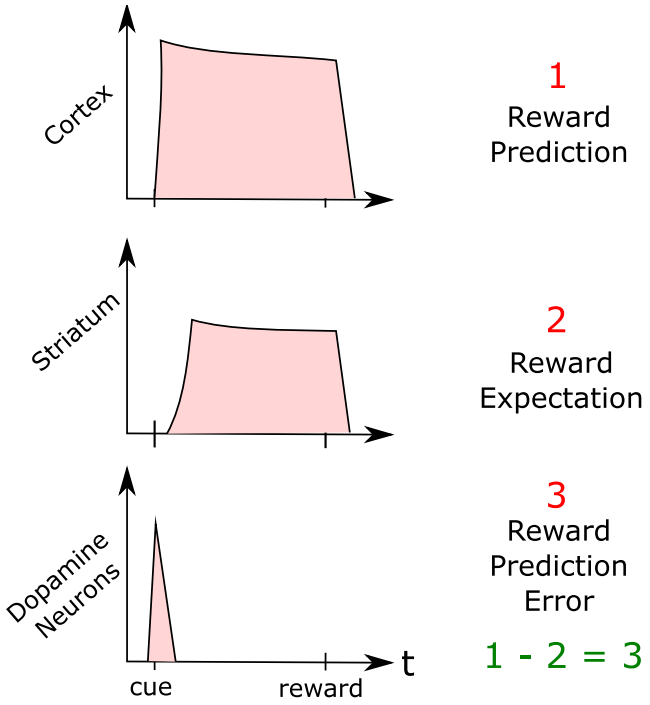

Figure 5.7: Our model proposes that the reward prediction error apparently signalled by dopaminergic nuclei is the result of two processes — a reward prediction, which comes from the cortex, and a delayed reward expectation which comes from inhibitory neurons in the striatum. I propose that these two processes are backpropagated in concert, so that at the end of the learning process, the sum of the excitation and inhibition results in a reward prediction error signal like the one shown above.

of inhibitory synapses on dopaminergic neurons, because the effect of expectation is to lower the likelihood of dopamine neurons from firing. Dopamine neurons receive inhibitory projections from many sources, including autoinhibition from dopamine released from the dendrites (reviewed in section 2.7.1). One major sources of inhibitory input is the striatum, which is composed primarily of gabaergic medium spiny neurons, and is uniquely placed to integrate multi-modal sensory input due to the massive convergence of cortical projections on the medium spiny neurons (Voorn et al., 2004). The striatum receives much of its input from the cortex, and is known to carry information about reward contingencies in reinforcement learning tasks (Haber et al., 2006). I propose that neurons in the striatum take information about reward-contingent sensory stimuli and integrate it to provide an inhibition of the dopaminergic neurons. *It is this inhibition that constitutes our expectation of reward.*

Another key factor that makes the striatum a suitable source for our proposed reward expectation is that it represents an indirect pathway by which sensory stimuli represented in the cortex can inhibit dopamine neurons. The indirect pathway would result in reward expectation arriving slightly later than the reward prediction (relative to the timing of the stimulus), which might prove quite a useful property — the onset of a predictive cue would trigger a strong, direct prediction form the cortex, and this would be followed by a slower, indirect inhibition from the striatum. This inhibition would serve to prevent dopamine neurons from phasically firing during the delay period, despite the sustained activity in the cortical populations. This process is illustrated in Figure 5.7.

One advantage of using expectation as the mechanism for backpropagation of reward prediction error is that it may also explain other well known reinforcement learning observations as a side-effect. Phenomena such as the dip in dopaminergic firing when an expected reward does not arrive, a lack of dopamine response to "blocked" stimuli, and a negative followed by positive reward prediction error when a reward is delayed, are all compatible with this mechanism (Schultz, 2007).

My aim in this section will be to demonstrate how the mechanism could work using a computational model.

5.8.1 Method

An updated schematic of the anatomy that the model represents is shown in Figure 5.8.

In this third and final computational model, inhibition is modelled by taking the

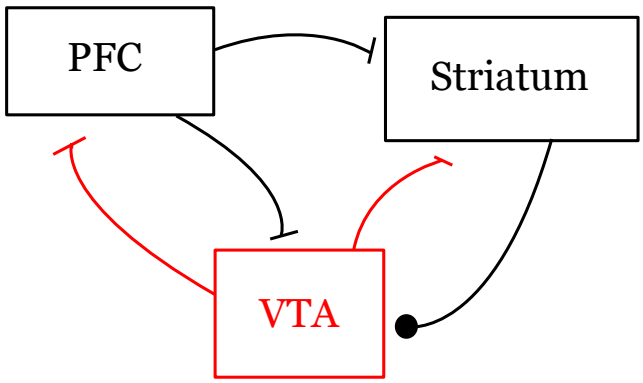

Figure 5.8: The proposed anatomy of the reinforcement learning circuit. The VTA releases dopamine according to the combindation of excitatory input from the cortex, and inhibitory input from the striatum. Populations in the striatum are driven by activity in the cortex, and so information about reward contingent stimuli arrives slightly later than direct information from the cortex. The subcortical sources of primary reward are not shown here.

activity from the cortical populations at the previous timestep and multiplying it by -0.6. The inhibition contributed by each population is subtracted from the excitation, and if the remaining activity of any one of the populations is greater than the threshold, a phasic spike of the dopamine neurons will occur. The value of 0.6 was chosen so that strong inhibition will lower the activity of dopamine neurons below the threshold for phasic spikes (0.4). Therefore a strong expectation of reward can be enough to prevent phasic dopamine release.

A reinforcement learning task is learned more quickly if there is a shorter time gap between the cue and the reward, or if the are many intervening cues between the first predictive cue and the reward. To help the model learn more quickly 4 serially ordered cues were used, with each one indicating that a reward will occur with 100% probability. If the model is working correctly, the peak reward prediction error and subsequent dopamine release should gradually backpropagate from the last cue to the first.

In our network the recurrence rates are bounded at 1, whilst a separate inhibitory weight is bounded at 2. This is to allow the inhibition to catch and eventually silence excitatory stimulation as long as learning continues.

The simulation was run in the same way as the previous two phases of the model, only for this model the simulation was run for 18 alternating rewarded and unrewarded trials.

5.8.2 Results

The results from this final phase of modelling are shown in Figures 5.9 and 5.10. The figures show that the mechanisms I have proposed can in principle lead to the back-propagation of reward prediction error. I have deliberately chosen simple mechanisms which have some grounding in the data, and showed how the interaction of these mechanisms can potentially account for the highly complex goal-driven behaviour we observe in animals.

5.9 Discussion

The model that has been constructed in this chapter successfully implements a reinforcement learning system inspired by the anatomy of the frontal cortex, dopaminergic nuclei, and the striatum. The intention of this model was to show how in principle these anatomical regions might combine to display the kind of complex goal-driven behaviour we observe in animals. It should be clear however, that the parameters of the model that was constructed were parameters that were chosen to make the model work, and are not based upon empirical values. As a result the model is quite inflexible, and is only capable of solving these toy problems in the limited parameter range within which it was designed to work. It is not my intention to extend the model to try to explain more data and make it more robust, but to highlight the kind of questions that could be examined experimentally to provide support or refute the validity of this model as a conceptual framework. As a computational modeller I want to avoid the temptation of making the model more complex to account for a greater range of behaviours - this could be viewed as a wasted effort unless the basic assumptions of the model can be justified experimentally. To make it explicit, the assumptions that were most critical for this model are:

1. That the effect of phasic dopamine release is to selectively enhance the persistent activity of neural populations active at the time of reward

2. That the VTA is subject to feed-forward inhibition that arrives slightly delayed relative to the stimulus induced excitation.

3. And that it is this interaction between prediction and expectation that leads to the back propagation of reward prediction error.

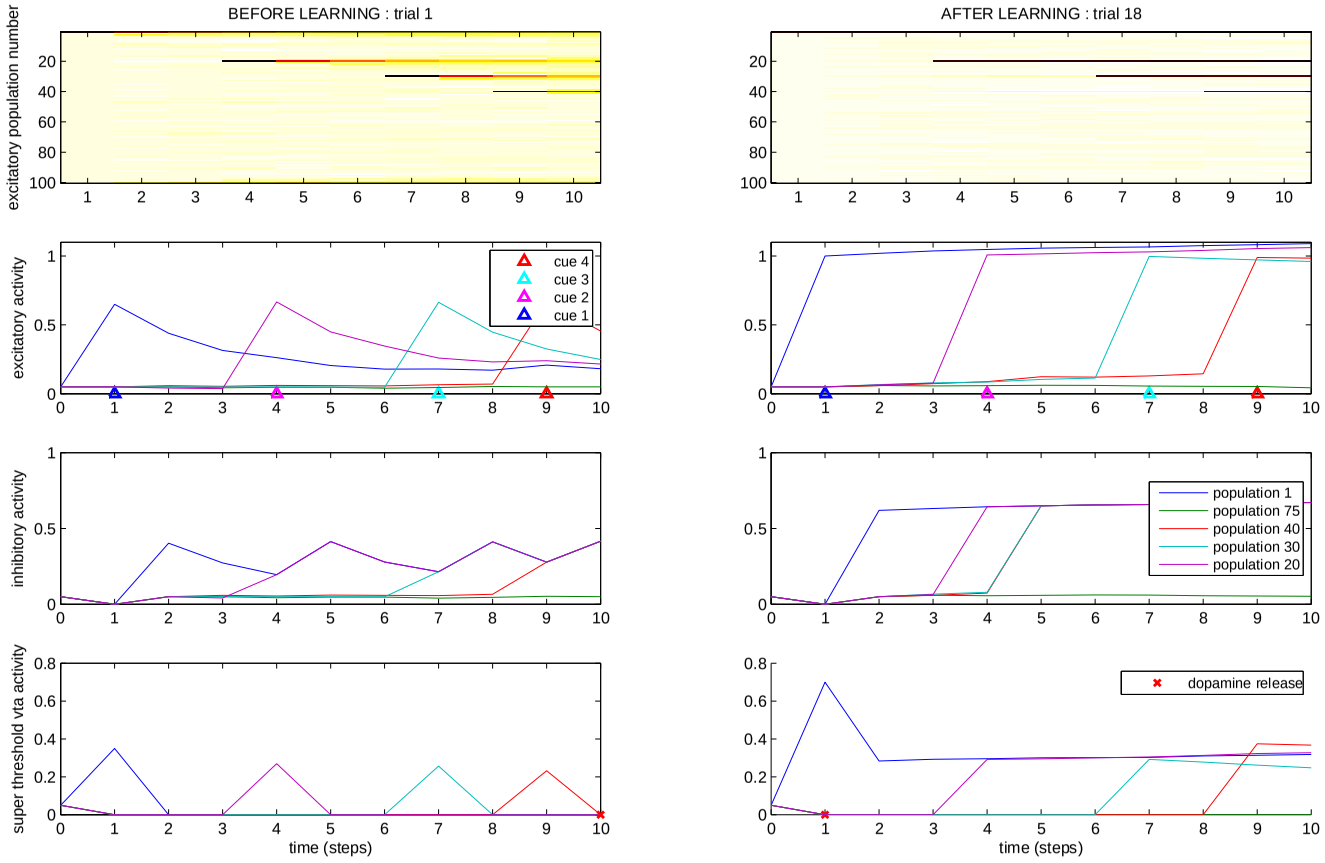

Figure 5.9: The four graphs on the left show population activity before learning has taken place, whilst the four graphs on the right show activity at the end of the learning process. The two graphs at the top show the activity of the whole cortical network as a heatmap. Before learning the cue-related activity dissipates quickly, but after learning the activity is maintained as indicated by the red traces on the right heatmap. The bottom left plot shows that before learning, each salient stimulus is capable of causing the release of some dopamine. After learning however, the bottom right plot indicates that persistent stimulus-related activity is counterbalanced by the persistent inhibitory activity that is coming from the striatum. The end result is that dopaminergic activity peaks at the onset of the first cue, and subsequent activity does not drive the dopamine neurons to fire as quickly. If the model was working perfectly we would expect the dopaminergic activity to return nearer to baseline (zero) after the initial cue.

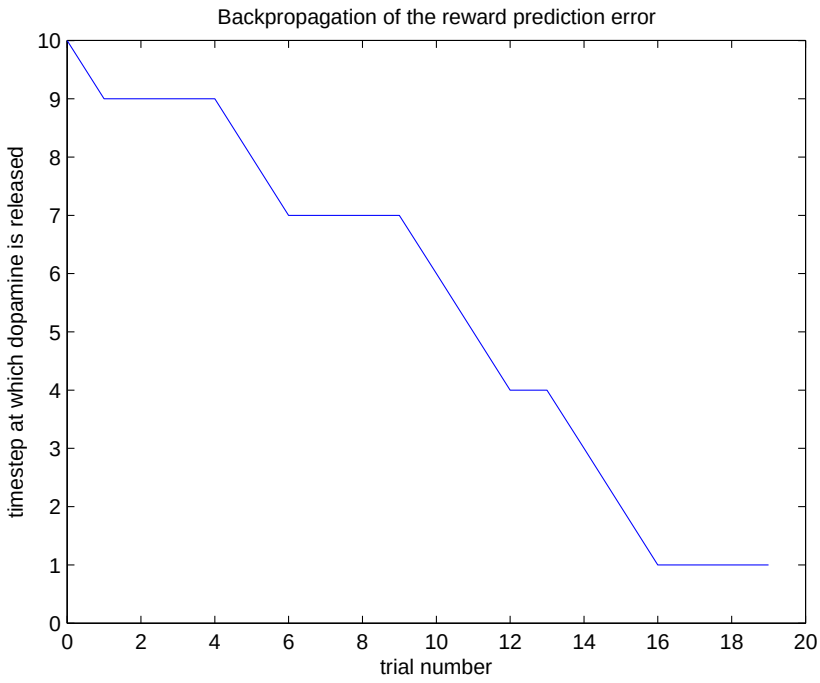

Figure 5.10: How the time of phasic dopamine release changes as learning progresses. This illustrates the process of backpropagation of reward prediction error with each repeated trial.

Without any observations to support the first assumption my model of dopamine-modulated plasticity would not be justified, and without any observations of the second my model of reinforcement learning would not function. Aside from these major points, there are a few other smaller questions to clear up. The question of how a combination of excitation and inhibition of the VTA influence the spiking behaviour of the VTA is a complex one, and one for which there is sparse empirical data. We have made some quite simple assumptions about a fixed firing threshold, and a simple summation of excitation and inhibition — the real picture is likely to be much more complex. The VTA receives glutamatergic, gabaergic, cholinergic, and dopaminergic input, and it is not know how these inputs overlap spatially, and how they interact to influence membrane dynamics. If there is some spatial segregation of inputs, or even topographic distribution of synapses this would have a profound influence on behaviour, and the types of rewarded tasks that can be learned.

One other reason why I have chosen to not develop this model further is that I have accounted for a complex behaviour while knowingly leaving out parts of the neuroanatomy known to be involved. Although there has been a recent trend to talk about reinforcement learning as a product of a dopaminergic learning circuit, it is clear that many other neuromodulatory and endocrine systems such as serotonin (Nakamura et al., 2008), noradrenaline (Aston-Jones and Cohen, 2005), acetylecholine (Kobayashi and Okada, 2007), and hypocretin (Korotkova et al., 2003), are all involved. We should be wary of using too little anatomy to explain too much behaviour.

Despite the limitations of this model there are a couple of positive outcomes of this model. I have shown that *working memory can in principle be seen as a subset of reinforcement learning*. The model has also made a clear prediction that *it is the relative timing of excitatory and inhibitory input that causes the backpropagation of reward prediction error*. This is a novel prediction that is potentially testable by experiment.

Since developing this model I have came across another model in the literature that appears to solve the same problem in a similar way (Ludvig et al., 2008). The model in this paper solves the problem by implementing the Temporal Differences algorithm with a *diffuse temporal representation* of the stimuli. In this model the Temporal Differences algorithm solves the problem of backpropagating the reward signal, whilst the diffuse temporal representation of cue stimuli is effectively equivalent to diffusion of the stimulus in the network in my model.

Although the two models appear to solve the same problem, both models have different strengths. (Ludvig et al., 2008) base their model on the Temporal Differences

algorithm, which is well understood and gives their model a sound theoretical foundation. *The model I have developed in this chapter is aimed more at describing the process of reinforcement learning in terms of the neurophysiology*, and so is possibly easier to use to interpret empirical data.

5.9.1 A review of the modelling approach

In this chapter I have tried to avoid tackling the question of what is really happening in the brain. This proved an advantage because it made the process of modelling easier — by side-stepping questions about how models relates to the biology, I was able to quickly set up a model that captures how the problem might be solved at a conceptual level. However, the corollary of this is that it is difficult, or perhaps ill founded to suggest that the model really represents how the biology solves the problem, rather it is a conceptual model which may or may not be consistent with the data. The aim with this kind of conceptual model is not to discover how the neurophysiology works, but merely to provide a framework for interpreting results, and planning future experiments or clinical interventions.

In some sense this approach has been successful in that I have been able to create a model which can in principle solve the problem, and I can relate this conceptual model to the brain at an anatomical level. However, the model could easily be criticised as a box-and-arrow diagram given anatomical dressing, and to some extent this criticism would be justified. But then one might argue that many of the best models are also box-and-arrow diagrams — sometimes it is useful to abstract away from the biological complexity in order to arrive at a model that is conceptually easy to understand, and therefore useful even if quantitatively inaccurate (Shouval et al. (2002) model of calcium-based plasticity could be seen as an example of this).

We have repeatedly stated that this approach to modelling is different in the sense that it is not interested in how the physiology really works, rather it aims to provide a conceptual model which can be related to anatomical and neurophysiological data. But is this really a novel strategy?

Taken to extremes we might wonder what value this approach has — if it makes no claims to have a 1-1 correspondence with the biology, then what do the components of the model correspond to? Either the model is shaped somehow by biological data, and therefore can be falsified by it, or it is a fantasy based upon preconceived ideas about how the brain ought to work. Some would argue that all such conceptual models have

their roots in our preconceptions of how mental processes occur, and it is only later that we look for evidence to support them. Perhaps the value of this approach is that it is quite explicit about the fact that it is constructed from a preconceived model of cognition (reinforcement learning), and therefore should not be confused with a model which claims to be saying something about how the brain really works, which as we have concluded in the previous chapters, is a sticky topic.

Although this third approach has offered a different method for answering the question that motivates the overall thesis, I don't believe it has squared the circle of finding a problem-free method by which modelling can participate in neuroscience. As a scientist it is distinctly unsatisfying to propose a model of reinforcement learning without also being able to suggest it is *the way* in which the brain solves the problem. If science is the process of understanding and predicting an objective world, then the approach we have described in this chapter could not be classed as scientific.

5.10 Conclusion

The model has shown how simple assumptions about the effects of dopamine and the anatomy of the basal ganglia may be enough to explain the phenomena observed during reinforcement learning tasks. By focussing upon processes in the brain which occur at a similar timescale to the behaviour itself, I was able to construct a model that is conceptually simple, but can still explain the key neurophysiological phenomena that have been observed in experiments. The units in the model are populations of neurons, and so the model is not committed to any particular mechanisms at the level of individual neurons (such as STDP), as is often the case in computational neuroscience.

However, there are issues that come with working at this level of detail in that it is difficult to support or refute this "interpretation" because it does not makes strong claims about the biology.

These issues can serve as both and advantage and a disadvantage. This model was an attempt to provide an alternative to the two models that went before it, and although it has succeeded in some sense, it falls short in others. If the model is focussed towards interpreting data rather than modelling and predictive objective facts then it is not scientific, however useful it may be conceptually.

Chapter 6

Discussion

6.1 Review of the modelling work

Before I begin the discussion I will give a quick summary of what was found in the three chapters of modelling in this thesis

6.1.1 Model 1 : Reinforcement learning by dopamine modulated spike-timing dependent plasticity

Model 1 was based upon an existing model of classical conditioning. Upon analysing it, I found that the model showed behaviour that was incompatible with it being a model of classical conditioning. Fixing the problems in the model effectively required rebuilding a new model from scratch, so I decided against doing this until a later stage (see Model 3). By examining what went wrong with the model I concluded that if a model is to produce falsifiable predictions then it ought to emerge from physiological data rather than theoretical models.

6.1.2 Model 2 : The effect of dopamine upon synaptic plasticity

In this chapter I attempted to construct an empirically-derived model that looked at the effect of dopamine on learning and whether or not the effects support the idea that it is implementing reinforcement learning. One result of the model was that it suggested that the synaptic effects of dopamine modulation would have a more significant effect upon synaptic plasticity than changes to intrinsic excitability.

In my conclusions I suggested that despite my best intentions the data in the model was composed from so many sources and theoretical models that it was unrealistic

to expect the model to make accurate quantitative predictions that generalise to other regions of the brain. I found that even when sticking close to the data there are many implicit and explicit theoretical assumptions that creep into the model and make it difficult to relate to experiments. The experience showed me that *the closer one gets to quantifiable empirical models, the further one gets from the questions that motivated the study in the first place*

My personal conclusion was that although this model could be improved — such as by improving the neuron model - some of the core problems found will not go away. Abstract models don't relate to the data, and are difficult to falsify. Purely empirical models don't exist — while you can get more detailed models, as you get more accurate you lose the ability to generalise the model to other systems.

6.1.3 Model 3 : A conceptual model of physiological basis of reinforcement learning

In this chapter my aim was to construct a conceptually valuable, rather than focussing upon making a model that was verifiable. The model was capable of solving the problem, and demonstrated a way of conceptualising how reinforcement learning may occur in the brain. However, the process was not satisfying on a scientific level, as it made no claims to be being an accurate representation of processes going on in the brain.

6.2 Discussion

This section would normally be focussed upon discussing the results of the models, but for me the most interesting outcomes of these chapters was not the data produced by the models, but what was learned in the process of modelling. Although the models did provide some interesting results (and these are discussed in each modelling chapter), on the whole the models failed to answer satisfactorily the questions I set out with at the start of this thesis.

The primary aim of this thesis was to use computational modelling to determine whether or not dopamine forms part of a reinforcement learning circuit in the brain. I have tried to answer this question in three different ways, and on each occasion I have run into methodological problems that prevented me from supporting or refuting the hypothesis.

When working with abstract models I found it difficult or impossible to relate them to the biology, and when working with biological-based models I found it difficult to relate them to a systems level theory.

My conclusion of this work is that it is not possible to develop computational or theoretical models of psychological phenomena that are both based on low-level empirical data, and are capable of making accurate predictions of new phenomena at the same scale. Making models that relate to psychological concepts requires that they be describable at a high level, whilst making them empirically justifiable requires that they be in agreement with a lot of low-level observations. As yet there are no cross-level theories of this kind in neuroscience, so in a sense it should not be surprising that my attempts failed.

That such cross-level models are *possible* is an article of faith in neuroscience, but my experience in this thesis has led me to no longer agree with this view. Physics, which is often held up as a paradigmatic example of successful science, has never developed successful cross-level theories. Despite the success of quantum mechanics and general relativity in their own domain, there have never been theories that unite both the large and small.

6.2.1 Neuromodulation and reinforcement learning as a paradigm

The discussions so far relate to the first question in my thesis as to whether or not dopamine is part of a reinforcement learning circuit in the brain. The second question in my thesis was whether or not reinforcement learning driven by neuromodulation was a suitable model for understanding other phenomena in neuroscience.

It was thought originally that the case of dopamine would provide strong support for the idea that we can understand much of behaviour in terms of a reinforcement learning circuit. The approach was unsuccessful and so the conclusions of my modelling have been that it is not possible to prove or disprove such a cross level model. However, I also concluded that not being able to prove the model true is not necessarily a problem, as theoretical models can have uses irrespective of their truth — this was discussed in more detail in chapter 5.

Like the idea of neuromodulatory driven reinforcement learning, the approach of understanding neuroscience in terms of spikes and computation is also something which cannot be proved or disproved. Perhaps we should see these theoretical frameworks not as something which can be true or false, but something which may or may

not be useful depending upon the context.

6.3 Support for the neuromodulation and reinforcement learning model

Despite my own personal conclusions regarding the mechanistic basis of the dopaminergic reinforcement learning hypothesis, there are many people who make use of this model, and believe in its validity without the need for an underlying mechanistic model justifying it with low level empirical data. To date, the original publication proposing the dopaminergic reinforcement learning hypothesis (Schultz et al., 1997) has been cited 2035 times, many of which by researchers who make use of the hypothesis despite its uncertain status. Why is it that some people do not feel the need for the hypothesis to be linked with an underlying model?

Different researchers make use of this model for different reasons, and not all of them are interested in whether or not the model is empirically justified. As I have already stated, there may be clinical situations where such a model is useful independent of how true it is. Indeed for the more pragmatic, usefulness is an adequate measure of truth (Hacking, 1983). But what other motives are there for people to believe in this model as a reasonable way of relating brain and behaviour, despite the apparent lack of evidence?

1. It provides a simple way of understanding complex human behaviour as a rational process

2. It offers an alternative viewpoint to the more common computational paradigm. In particular it *offers a view of human behaviour as a product of our desires (drives), rather than our desires being a product of computation*.

3. Reinforcement learning is fundamentally about causality — learning what cues cause what rewards. As such it is restating basic scientific principles, and offers researchers a way to describe human behaviour in the scientific image.

4. Phasic dopamine release correlates with the reward prediction error signal in reinforcement learning, and *for some this is enough evidence* to believe that dopamine is sending this signal.

If we are interested in the reasons why people have adopted reinforcement learning and neuromodulation as a basis for explaining the relationship between brain and behaviour, then it is worth looking at the historical background of this model. As explained in section 2.3, reinforcement learning and the temporal differences algorithm grew out of work by engineers and mathmaticians who wanted to characterise the optimal way to make decisions. A correlate of this abstract framework was later discovered in the brain. Does this tell us that:

(a) these engineers and mathmeticians had good insight into how people make decisions.

(b) these engineers and mathmaticians formalised their own ideas about what kind of decisions are optimal. Scientists also share those ideas, and so when they found similar processes in the brain, and were able to collect and frame the evidence in terms of the theory.

My aim is not to suggest that scientists have consciously created both the theory and the evidence, but rather to highlight the sociological factors that can lead to a new theory becoming popular. I believe that reinforcement learning is not a model of behaviour which is uniquely supported by the evidence, but a subsconscious reflection of its authors views on human behaviour.

Much of the data that is used to support reinforcement learning as a model of behaviour is evidence that is chosen from a pool of contradictory observations. The evidence which is chosen is that which will fit a pre-existing model of how humans behave — a model that mirrors the way its authors behave.

If this is true there ought to be lots of observations which contradict the dopaminergic reinforcement learning hypothesis, and there are. As I progressed in my thesis I became more and more aware of evidence that appeared to contradict the hypothesis. Strangely this data does not get nearly as much attention as data that supports the model. While I still believe that the dopaminergic reinforcement learning model is valuable, I think it is important that we are aware of evidence that contradicts it. I will highlight some of that evidence in the next section.

6.4 Evidence against the neuromodulation and reinforcement learning model

There are good reasons for emphasising that the hypothesis is not uniquely supported by the biological data. Despite the very strong case offered by the correlation between the activity of dopamine neurons and a reward prediction error signal, there are cases where these two variables cease to correlate, indicating that the hypothesis may only be true in certain situations.

During the course of my research I found several experimental reports with evidence that did not fit with the view that dopamine is signalling reward prediction error.

6.4.1 Dopamine neurons fire phasically during sleep

One particular finding is that dopamine neurons fire phasically during sleep, when there is no reward prediction error in any meaningful sense (Dahan et al., 2006). If dopamine neurons cease to signal reward prediction error during sleep then it is worth asking what happens at the onset and offset of sleep to the reward system? By understanding how dopamine neurons change from their daytime to their night time behaviour we might gain some insight into the function of dopamine neurons in a broader context.

However, at present data regarding the firing of dopamine neurons during sleep is sparse.

6.4.2 Dopamine is also released by noradrenaline neurons

Another interesting result is that noradrenaline neurons also synthesise and release dopamine (Devoto et al., 2005). Dopamine is a chemical precursor of noradrenaline, and so significant quantities of dopamine are synthesised by noradrenaline neurons during the production of noradrenaline. Although this finding does not contradict the observation that dopamine neurons signal reward prediction error, it does muddy the waters a little with regard to the proposed function of dopamine. It is generally thought that dopamine itself is the means by which dopamine neurons signal reward prediction error to the neurons within their axon terminal field. However, if dopamine can be coreleased by noradrenaline neurons in a way which does not correspond to reward prediction error it does indicate that there is more to dopamine than signalling reward prediction error.

6.4.3 Not all behaviour can be understood in terms of reward

This point may be as much a philosophical one as an empirical one, but there are many who would argue that it is impossible to describe some human behaviour in terms of maximising a reward function. This debate has been ongoing for many years (Dawkins, 1990) (Okasha, 2009), and its lack of resolution indicates that this model of behaviour will also fall short for some when it comes to describing altruistic behaviour.

6.4.4 Dopamine may be too slow for reward learning

The initial observation that drove this thesis was that dopamine appeared to be involved in learning the causes of reward, but recent research has called that into question. Electrophysiological studies of the effect of dopamine modulation have shown that by itself, dopamine modulation is too slow acting to capture trace information about the reward, and therefore cannot be the only factor in reward learning. Rewarding stimuli in conditioning experiments may only be present for a matter of seconds, but the major effects of dopamine occur in the minutes following phasic release (Lapish et al., 2007). It has been suggested that the proposed reinforcement learning circuit may in fact be dependent upon the co-release of glutamate from dopaminergic neurons (Lavin et al., 2005). This is consistent with the observation that the initial effect in the cortex of phasic spiking in dopamine neurons is glutamate-dependent. As we learn more about reinforcement learning circuits in the brain we may find that they are as much dependent upon glutamate release as they are upon dopamine.

6.5 An alternative view of dopamine modulation

Despite the popularity of the hypothesis that dopamine signals reward prediction error, there are some who argue that this is not the whole picture. Redgrave et al. (2008) argue that the ability of dopamine neurons to signal non-rewarding novel stimuli overlaps with their ability to signal reward-related stimuli, such that in natural environments what dopamine neurons respond to is an amalgamation of both signals.

The authors also argue that phasic dopamine responses are too fast to come from sensory processing of stimuli, and can only occur as part of a pre-attentive circuit. They propose that collicular neurons detect early visual changes, and if there is no inhibitory input coming in to the dopaminergic nuclei, this information is deemed to be unpredicted, and therefore worthy of further learning.

They authors suggest that dopamine neurons do not signal reward prediction error, but act as *a reinforcement signal for the organism to learn what part of its own behaviour caused an unexpected event*. According to this view *it is agency which defines reward*, and not the other way round.

6.6 Is dopamine a unitary entity?

Since the dopaminergic reinforcement learning hypothesis was put forward, other researchers have stepped in to propose new behavioural functions for additional neurotransmitters (Yu and Dayan, 2005), (Yu and Dayan, 2005), (Dayan and Huys, 2008), (Aston-Jones and Cohen, 2005), (Redgrave et al., 2008). In section 1.6 I argued that this was part of a growing trend. While this approach is promising, it is also slightly misleading. Although neuromodulators like dopamine are relatively simple chemical signals, it is oversimplistic to treat them as unitary entities and assign them a single function. Basing models around a neuromodulator may appear to provide a simple explanation of behaviour, but in reality what is really does is hide the complexity elsewhere - in the release, the binding, and the receptor dynamics of the modulator.

At present, our knowledge about the function of dopamine comes largely from studies of the firing of dopamine neurons, studies of the systemic effects of dopamine release, or studies of the acute effects of dopamine in isolated preparations. While these experiments can tell us what processes dopamine has some influence in, none of these situations really give us enough experimental control to determine what dopamine does and assign it a single function. Until the day we can perform these ideal experiments it is worth thinking of dopamine as having multiple roles, because it is possible that *when the effects of dopamine are better understood it may be found to fulfilling many complex and contradictory roles through its family of receptors and binding sites*.

6.7 Summary

The models developed in this thesis were unable to conclusively support or contradict the hypothesis I set out to examine. However, the process of constructing and evaluating these models did lead me to rethink my methodology, which in the end I feel, has proved more valuable than the quantitative predictions produced by the models. My conclusion from this work is that at present it is not possible to provide a mechanistic

model of how or why dopamine neurons signal reward prediction error. My reasons for reaching this conclusion is that *there is simply too much relevant complexity in the system to produce simplified models and expect them to provide accurate quantitative predictions*.

Although my models have failed to support or contradict the hypothesis, my experience in constructing these models has led me to view the overall neuromodulation and reinforcement learning paradigm in a different way. I do not believe that this model is born out of the empirical data, but I also do not believe this is a barrier to using the paradigm in interesting and useful new ways.

Developing these models and trying to answer the questions in this thesis has led me to conclude that it is not possible to develop cross-level theoretical models of what is really happening in the brain, in fact I think the question is ill-posed. However, I do believe that it is possible to develop theoretical models that are useful, and whose value is not based upon their accuracy.

Much like the paradigm of computation and spikes, neuromodulation and reinforcement learning is not something which is true or false, but merely a framework for interpreting biological data. As we begin to collect more data that allows us to relate neuromodulation to behaviour, this framework is likely to prove ever more useful.

Chapter 7

Putting dopamine and reward in context

7.1 Novel contribution

In this chapter I review the many transmitter systems that modulate dopamine neurons, and use this as a basis for exploring the context of dopamine modulation and reward. I suggest that defining dopamine function only in terms of reward obscures its relationship with other evolutionary critical bodily processes. In particular I examine the function of dopamine in the context of stress, and propose a way in which dopamine modulation can be seen as part of an extended stress response. I also relate dopamine modulation to immune processes and suggest that there is value in looking at dopamine as a signal of agency as much as reward.

7.2 Précis

Throughout this thesis I have described dopamine purely as a signal of reward prediction error. While there is strong evidence that dopamine is involved in reward processing, it is also involved in many other processes, and our current fascination with the reward system serves to obscure these roles. As long as we choose to define dopamine neuron activation with respect to reward, it will remain difficult to relate dopamine to other bodily processes that are not shaped by reward.

The aim of this chapter is to provide an alternative view on dopamine modulation. The outcome of the modelling work in this thesis has been to show that the idea that dopamine modulation is implementing a reinforcement learning circuit in the brain is

just one of many possible interpretations of the function of dopamine. This interpretation is useful if we are interested in the domain of goal-directed behaviour, but is not useful if we are interested in behaviour that is not driven by a desire for reward. There are plenty of papers that attempt to show how dopamine modulation implements some additional facet of reward-related behaviour, but there are few which attempt to relate dopamine modulation to other non goal-directed processes.

In this chapter I review the other transmitter systems which modulate dopamine release, and use this as a way of introducing other bodily processes which shape reward-driven behaviour. In particular I will look at the function of dopamine in the context of stress and immune responses.

7.3 Introduction

The observations of Schultz in the 1980s that dopamine neurons fire phasically when an animal receives an unexpected reward has led to the widespread view that dopamine plays a major role in the brains reward circuitry. For a review see (Schultz, 1998).

Following this discovery, it was realised that the pattern of firing of dopamine neurons is reminiscent of the reward prediction error signal described by the temporal differences algorithm (Schultz et al., 1997). This apparent correlation between the actions of neurons in the midbrain and the workings of an abstract model of learning has led to the view that the function of dopamine circuitry in the brain is to implement reinforcement learning - a hypothesis I will refer to as dopaminergic reinforcement learning. This view of behaviour as driven by the action of neuromodulators rather than ensembles of neurons, was radically new, and over the last 10 years has developed into a new paradigm. Although originally based around the actions of dopamine, theorists were quick to propose roles for other neuromodulators (Doya, 2002). It has since been suggested that acetylcholine signals expected uncertainty (Yu and Dayan, 2005), noradrenaline signals unexpected uncertainty (Yu and Dayan, 2005), serotonin mediates behavioural inhibition (Dayan and Huys, 2008), and noradrenaline mediates the exploration:exploitation trade-off (Aston-Jones and Cohen, 2005). This view of behaviour has proven popular because it has allowed theorists to talk about how motivational states, as signalled by neurotransmitters, could come to shape behaviour. This is in contrast to the more traditional models in computational neuroscience which have tended to explain behaviour in terms of neural coding based upon neuron action potentials.

But as we learn more about dopamine and neuromodulation, it is becoming clear that this view of behaviour whilst powerful, is not the whole picture. While the dopaminergic system is clearly involved in reward-related behaviour, it appears that it also plays a role in more fundamental behavioural circuits.

The aim of this chapter will be to explore these circuits, particularly those originating in the neuroendocrine and immune systems. I will look at how these systems modulate dopamine release, and describe a conceptual framework with which we can understand how dopamine can play a role in not only reward-related behaviour, but also behaviour which serves the more fundamental needs of the body.

What reasons do we have to believe that dopamine modulation and reward-driven behaviour is shaped by more fundamental forces? In sections 7.3.1 and 7.3.2 I will look at biological data and historical evidence that indicates that we need to see dopamine modulation and reward-driven behaviour in the context of more evolutionary critical processes.

7.3.1 The biological case

The dopaminergic reinforcement learning model as described by Doya (2002) is powerful; taken literally it suggests that *behaviour is a result of reward maximisation*, and within this reinforcement learning circuit dopamine signals reward prediction error and reward value (Tobler et al., 2005).

The power of this model means that it can account for a large range of motivated behaviour, but so far, few of the explanations of goal-directed behaviour provided by this model relate to our explanations of other fundamental bodily processes. As yet the dopaminergic reinforcement learning model has not been integrated with models of basic bodily functions such as hunger, thirst, sex, and immunity, and how they come to shape what we think of as rewarding.

In this section I will attempt to *explore the context of dopamine and reward by looking at the transmitters and peptides that modulate this proposed reinforcement learning circuit*. The aim is that by looking at the signals which shape dopamine release, we can gain a better understanding of the function of dopamine in the context of other more fundamental behavioural circuits.

The neurotransmitter systems in the midbrain are highly interconnected, and the actions and concentrations of one neurotransmitter will inevitably affect others even if there is no direct interaction. For simplicity I will focus only on the neurotransmitters

which interact directly with the substantia nigra and ventral tegmental area.

A list of the neurotransmitter receptors known to be expressed on these dopaminergic neurons is shown in Table 7.1. When any of the receptors are stimulated, the electrophysiological properties of the dopamine neurons change. This has a knock-on effect on the firing patterns of dopamine neurons, which in turn affect their response to a conditioned stimulus.

To give an example of how other neurotransmitter systems can provide a context to reward-related behaviour, we can look at the actions of the neuropeptide orexin (also known as hypocretin). Orexin is produced in the lateral hypothalamus and perifornical area, and is involved in the regulation of feeding and wakefulness (Harris et al., 2005), (Willie et al., 2001). Orexin producing neurons contain projections to the substantia nigra (amongst many other midbrain nuclei), where there are orexin receptors (Peyron et al., 1998), (Korotkova et al., 2003).

In vitro data indicates that orexin increases the firing rate of dopaminergic neurons in the ventral tegmental area (Korotkova et al., 2003). Taken together these lines of evidence suggest that *orexin modulation of dopamine release provides a mechanism by which metabolic and neuroendocrine signals in the body can exert an influence on perception and action circuits in the cortex*.

This hierarchical modulation of cortical circuits via dopamine is one way in which endocrine circuits can influence global behaviour. But given that it has been claimed that dopamine alone signals reward prediction error, this implies that dopamine must serve as a single common pathway for any signal that is to influence behaviour that is shaped by reward prediction error.

The receptors listed in Table 7.1 indicates that there are a large number of transmitter systems modulating dopamine release - a sign that reward prediction error is subject to modulation by many other behavioural processes. Does this mean that modulation of dopamine release is the only way in which these other transmitter systems shape the perceived reward prediction error? Or do they act in parallel via many pathways?

In biology it is unusual to find a function supported by a single pathway. in contrast, the mechanisms we observe operate by multiple parallel pathways, making the functions they support robust to malfunction in just one pathway.

While dopamine does appear to be a good correlate of reward prediction error in controlled conditioning experiments, it seems unlikely that the brain is dependent upon this single pathway for all aspects of motivated behaviour. By suggesting that dopamine signals a general reward prediction error that alone provides the information

Reference	Region	Receptors found
Johnson and North (1992)	VTA	GABAA, GABAB, NMDA, D2
Barroso-Chinea et al. (2005)	VTA, SN	GDNF
Di Matteo et al. (2001)	VTA, SN	5HT2C, 5HT1B
Korotkova et al. (2003)	VTA	Orexin
Diana and Tepper (2002)	VTA, SN	Glycine, CCK, neurotensin, substance K, neurokinin NK3
White (1996)	VTA	M1, M2, and Nicotinic receptors
Margolis et al. (2006)	VTA	kappa-opiod receptors
Westerink et al. (1996)	VTA	GABAA, GABAB, NMDA, ACh
Wang et al. (2007)	VTA	CRF
Leshan et al. (2010)	VTA	leptin
Morimoto (1996)	VTA, SN	glucocorticoids
Kalivas (1993)	VTA	D2, GABAA, GABAB, neurotensin, CCK, EAA (NMDA and non-NMDA), mu-opiod, NK3, 5HT1B, 5HT1C, Nicotinic, M1
Ye et al. (2005)	VTA	IL-2

Table 7.1: The location and receptor types found on midbrain dopamine neurons

needed to learn goal-directed tasks, we are going against the grain and proposing a mechanism that does not sit well with what we know about biological systems.

If a reinforcement learning system does exist in the brain, it would be more robust and accurate if it used more than one transmitter system to estimate reward prediction error. The problem with this proposal is that if we extend the dopaminergic reinforcement learning model to incorporate the actions of many transmitters then we weaken the reductive power of the original dopamine-based model. It seems that by making the model more realistic, we make it less powerful.

Also noticable in Table 7.1 is the diversity of transmitter systems that provide context to dopamine modulation; classical neurotransmitters such as glutamate and GABA (Westerink et al., 1996), neuromodulators such as NMDA, and ACh (Westerink et al., 1996), neuroendocrine factors, such as CRH (Wang et al., 2007), and neurotensin (Kalivas, 1993), or cytokines such as IL2 (Ye et al., 2005) which are thought to play a role in immune and inflammatory responses.

The overlap of the reward system with neuroendocrine and immune systems is interesting because it suggests that reward circuits must also be an integrated part of the stress and immune responses, and *our models of reward-related behaviour must also be compatible with the conceptual framework of both of these fields.* In sections 7.5 and 7.6 I will look at how these circuits affect dopamine release. I will suggest ways in which the function of dopamine can be understood from within the conceptual framework of both these fields. Afterwards I will look at how well these viewpoints fit with the most recent evidence regarding the function of dopamine.

7.3.2 The historical case

The previous section outlined a biological case for why we ought to make an effort to understand the actions of dopamine and reward in the context of other behavioural systems. Alongside the biological case for wanting to base our models of behaviour on something more fundamental than a desire for reward, there are also historical reasons to believe that there is a limit to what reward-based models should be used to explain.

Questions about whether or not reward is a reasonable basis for models of behaviour can be traced back to the debates surrounding behaviourism in the 60s. Then, opponents of behaviourism contested that there were some behaviours that could not be explained as a result of reward or conditioning (Chomsky, 1959). Breland and Breland (1961) famously demonstrated out that that there are some 'instinctive' behaviours

which are difficult to override with reinforcement learning. They cite examples of animals choosing instinctual behaviours over those that were conditioned - chickens that were more interested in pecking at the CS than getting the reward, and pigs that would, over time, revert to rooting the CS rather than engaging in the rewarded behaviour.

At the time, these unconditioned behaviours were described as instinctive or innate, but it is possible that these behaviours are observed because they are the expression of drives and circuits not accessible by dopamine or reward. Because these drives are more fundamental than reward, no amount of experimental reward can change them.

The story of behaviourism tells us that models of behaviour based upon purely maximising reward have proved problematic before. To avoid repeating these mistakes we should seek to understand the biological context of reward learning - in doing so we might better understand the relationship of reward-seeking behaviour to other human behaviour.

7.3.3 Summary

Dopaminergic reinforcement learning is a fascinating model for explaining goal-directed behaviour. The development of this framework and its grounding in the biological data marks a qualitative shift from the models of behaviour that preceded it.

However, it is clear from both the biological data (Section 7.3.1), and historical precedents (Section 7.3.2), that there are caveats to this model. This chapter will attempt to review how reward and dopamine modulation interact with other behavioural systems. I will review how dopamine is modulated by stress and immune signals, and what this means for a model of behaviour based on reward with dopamine as a signal.

7.4 Dopamine modulation in context

7.5 Stress

The term “stress” was first used in biology by Selye (1955), to describe the body’s response to disturbances from homeostasis. In this chapter I will treat stress as synonymous with neuroendocrine stress, and assume that ACTH release is an accurate correlate of a stressful event having taken place.

During a stressful event, parvocellular neurons in the periventricular hypothalamus (PVN) release corticotropin-releasing hormone (CRH) into the anterior pituitary.

Within 15 seconds the presence of CRH stimulates the release of adreno-corticotropin hormone (ACTH) into circulation. A few minutes later ACTH in the adrenal cortex stimulates the release of corticosteroids. Corticosteroids are able to cross the blood-brain barrier into the brain where they inhibit CRH release, closing the circuit, and effectively switching off the stress response.

7.5.0.1 Modulation of dopamine release by stress

Stress provides context to reward processing most directly through the actions of CRH, ACTH, and corticosteroids.

In the central nervous system CRH acts primarily through volume transmission, and alongside the CRH released in the anterior pituitary, *stressful events result in an increase in CRH concentration in the VTA*. It is not known if the source of this is volume transmission from the PVN, direct projections, or co-transmission from other neurons (Wang et al., 2005). However, VTA application of CRF2R agonist was found to increase dopamine release in the VTA to 160 percent of the baseline concentration (Wang et al., 2007).

Whilst ACTH is found in the central nervous system, the ACTH that is released into the periphery following a stressful event does not re-enter the central nervous system.

In vitro corticosterone effects dopamine release via glucocorticoid and mineralocorticoid receptors (Rougé-Pont et al., 1999). In vivo, blocking corticosterone secretion with an adrenalectomy decreases stress-induced dopamine release in the nucleus accumbens (Rougé-Pont et al., 1998). Also peripherally administered corticosterone increases dopamine concentration in the nucleus accumbens, although the magnitude of the effect is dependent upon individual differences and the dark/light cycle (Piazza et al., 1996). Together these pieces of evidence demonstrate that the corticosteroids released following a stressful event are a potent modulator of dopamine release.

Corticosteroids and dopamine are both neuromodulators downstream of what we have defined as our correlate of stress - ACTH, and so in the same way as corticosteroids are thought of as part of the stress response, we can also think of dopamine modulation as part of the extended stress response. In sections 7.5.2 and 7.5.3 we will investigate what perspective it gives use to see dopamine function in the context of a stress response.

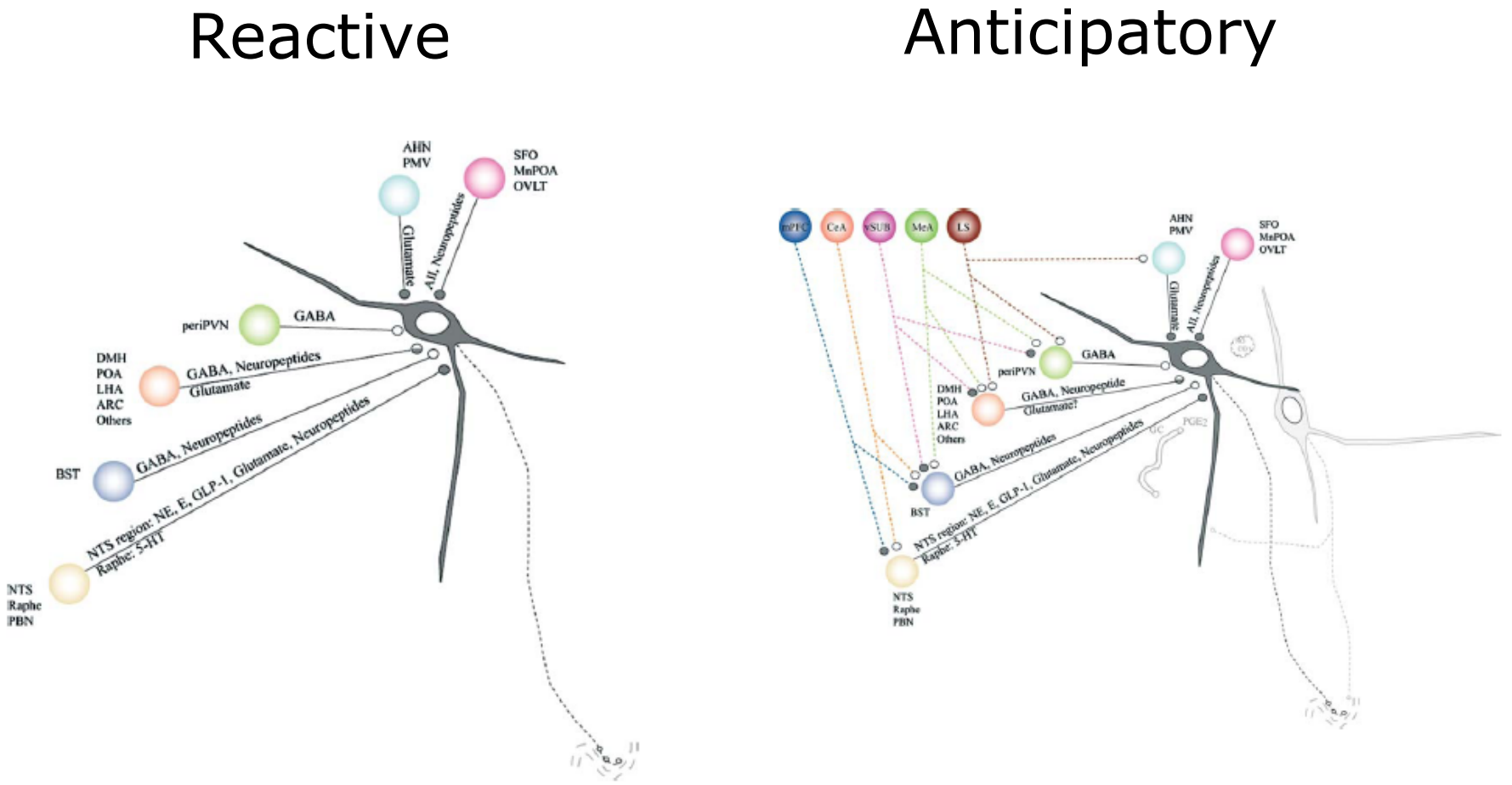

Figure 7.1: Regions that directly innervate the PVN relay sensory information from visceral afferents, nociceptors, and circumventricular organs and promote “reactive” responses to homeostatic challenges. Indirect inputs from limbic and cortical circuits are capable of activating these same cells in the absence of physiological challenges. These are described as “Anticipatory” stressors. Figure adapted from (Herman et al., 2003).

7.5.0.2 Modulation of stress by dopamine neurons

While the neurocorrelates of stress exert a potent effect on dopamine modulation, it is notable that *dopaminergic neurons in the ventral tegmental area and substantia nigra do not have strong reciprocal links to the PVN*. Regions that do modulate the PVN are shown in Figure 7.1. The effects that dopamine release has on stress are likely to be indirect, possibly through the effect of dopamine on neurons that do project to the PVN. For example, the prefrontal cortex, which is strongly influenced by dopamine modulation is known to indirectly modulate the PVN - for a review, see (Herman et al., 2003).

7.5.1 Dopamine and stress : Positive and negative feedback loops

The dynamics of the stress response and the dynamics of reinforcement learning suggest that we might expect to see interesting results when these two systems interact.

As described in section 7.5.0.1, CRH produces ACTH which inhibits CRH release and brings ACTH concentration back to basal levels. In this way the neuroendocrine stress system functions as a negative feedback loop, effectively guaranteeing the eventual shut-off of the stress response. However, the actions of dopamine, from the perspective of the reinforced behaviour, can be seen as a positive feedback loop. Unexpected rewards lead to an increase in dopamine concentration, which in turn promote the learning and repetition of the behaviour which lead to the reward.

This interaction between a positive and a negative feedback loop is potentially interesting. A negative feedback loop tends towards stability, whilst a positive feedback loop leads to runaway activity. However, when a positive and negative feedback loop interact there is a potential for complex and possibly chaotic behaviour.

The canonical example of this is the logistic map:

$$x_{n+1} = x_n - x_n^2 \tag{7.1}$$

an iterated map with a positive and negative feedback term, the logistic map is known to demonstrate chaotic oscillations (May, 1976). The properties of the logistic map were famously studied by Lorenz (1964) and Feigenbaum (1978). It is a matter for further investigations whether or not these chaotic dynamics also appear in the intersection of the stress and reward systems.

7.5.2 Dopamine as an anticipatory stress system

I have described in section 7.3 that dopamine is most often interpreted as a reward signal. However, the evidence that dopamine both modulates and is modulated by stress, indicates that *we ought to be able to understand reward as part of the stress response*.

In the stress literature it has been suggested that stressors can be distinguished into two categories, relating to the way in which stressful stimuli activate the PVN (Herman et al., 2003). 'Reactive' stressors relay an immediate homeostatic challenges directly to the PVN. Regions that directly innervate the PVN relay primary sensory information from the brainstem, visceral afferents, and nociceptors (see Figure 7.1). These inputs promote a reactive corticosteroid response to environmental stressors.

'Anticipatory' stressors on the other hand, indirectly regulate corticosteroid release under conditions where there is no immediate threat, but a homeostatic challenge is predicted. These anticipatory signals modulate the PVN via the projections of limbic and cortical structures (see Figure 7.1). The signals provided by these systems may come from innate programs, or can be learned through a process of conditioning.

As pointed out in section 7.5.0.2 dopamine neurons do not have a strong direct influence over CRH release from the PVN. However, dopamine's proposed role in reinforcement learning means that the actions of the midbrain dopamine system fits well with the idea that dopamine forms part of an anticipatory stress circuit. It is well established that dopamine predicts the appearance of one form of stressor - reward.

It may seem perverse to describe a reward as stressful, but a rewarding stimuli is a strong perturbation from homeostasis and is therefore a stressor. Selye himself recognised that stressors can have both positive and negative effects, and he described these aspects as "eustress" and "distress" (Selye, 1975).

If we think of the function of dopamine as to drive learning about the causes of stress, then it frees us up from the debate over whether dopamine is involved in signalling rewarding or aversive events (Frank and Surmeier, 2009). From the point of view of the stress system, both are stressors to be anticipated.

7.5.3 Dopamine as a generator of allostasis

Another concept in the stress literature which offers a fresh perspective on the function of dopamine is the notion of allostasis and allostatic load.

Allostasis is the process of maintaining stability through change, and is an impor-

tant counterpart to homeostasis - for a review see (McEwen, 1998). While we might naively think that homeostasis means attempting to keep the parameters of a system fixed, there are situations in an organisms lifespan where the set points of a system must change to maintain overall homeostasis.

A phenomenon that has been used as an example of allostasis is the fluctuations in blood pressure that occur during the sleep-wake cycle. When we get out of bed in the morning our blood pressure must rise to ensure a steady supply of blood to the brain. Although homeostatic systems tend to keep variables close to a fixed point, allostatic systems are necessary to enact change that will maintain overall homeostasis.

Allostatic change is an important part of homeostasis, but there is a cost of this adaptation, and this cost is known as *allostatic load*. Allostatic load occurs when there is a frequent activation of allostatic systems; when there is a failure to shut off allostatic activity after stress; or when there is an inadequate response of an allostatic system which has negative effects on overall homeostasis (McEwen, 1998). Allostasis has been described metaphorically as like the water used by firemen. If used sparingly it can put out the fire, but too much and the resulting water damage can be worse than the fire itself (Korte et al., 2005).

Allostasis and allostatic load may be important components of the stress system, but how does this relate to dopamine and reward?

The neuroendocrine stress system as we have described it is based upon negative feedback, and thus ensures that outside of stressful stimuli, corticosteroids are kept within stable levels. In this way we can view the neuroendocrine stress circuit as a homeostatic mechanism.

Dopamine however acts differently. By causing learning after the presentation of a reward, dopamine reinforces approach behaviour to a stressor. In this sense the significance of dopamine is that it forms an *allostatic circuit*, and rather than promoting behaviour that leads the organism to homeostasis, *it actually encourages the organism to engage in behaviour that takes it away from homeostasis*.

Allostasis systems are a focus of study because they underlie some of the most important behaviours in an organisms lifespan. Mating, rearing young, feeding, all take an organism away from homeostasis, and these behaviours are necessary for development.

Describing dopamine as an allostatic circuit makes explicit something that is rarely recognised when in the reward learning literature - that *reward-seeking behaviour carries with it some penalty*. Reward is generally thought of as a positive or neutral out-

come, whereas the concept of allostatic load implies that such behaviour must contribute wear and tear to the organism.

An over-reliance on reward can lead to a difficulty unlearning rewarded behaviours and can result in a reduced behavioural repertoire, both of which play a part in the development of obsessive-compulsive disorder. Taken to its extreme, purely reward-seeking behaviour has similarities with psychopathy. This behavioural phenotype correlates with a hyper-reactive mesolimbic dopaminergic system, and a increased risk of developing substance use problems (Buckholtz et al., 2010).

This perspective on reward and allostatic load is particularly interesting in the context of addiction. It has been suggested that addiction occurs when allostatic mechanisms take control of behaviour (Koob and Le Moal, 2001). It is known that dopaminergic systems are potentiated by stress, and that individuals subject to stress tend to engage in more reward-seeking behaviour irrespective of the damage it might cause.

7.5.4 Summary

From the perspective of stress, *dopamine modulation is an anticipatory stress circuit*, indirectly signalling predictions to the PVN of forecoming psychological and behavioural stressors. These stressors may be both rewarding and aversive.

Dopamine can also be thought of as an allostatic circuit, encouraging behaviour that takes us away from homeostasis. This fits well with the role of dopamine in feeding, sexual, and addictive behaviour. The concept of allostatic load suggests we should look more closely at the effect of reward-seeking behaviour, and the way in which it can impose negative consequences on the individual.

7.6 Immunity

7.6.1 Causal interactions between the immunity and reward

Immunity is another fundamental bodily process which has a strong influence over reward-seeking behaviour. During an immune response, peripheral nerves and pro-inflammatory cytokines send signals from the periphery which initiate what is known as sickness behaviour - a change from exploratory behaviour to a focus upon reducing energy loss and fighting infection (Maier and Watkins, 1998). Pro-inflammatory cytokines such as IL1 are produced in the periphery and central nervous system during an immune response (for a review see (Maier, 2003) and (Szelényi, 2001)), and when

administered centrally, these cytokines are known to reproduce many of the features of sickness behaviour such as decreased food intake, decreased sexual behaviour, and decreased locomotion (Maier, 2003). Conversely, blockade of IL1 receptors in the central nervous system is enough to blunt the sickness behaviour observed following administration of LPS (LipoPolySaccharide) (Bluthe et al., 1992).

At the behavioural level sickness behaviour expresses itself in many ways, amongst them is a reduction in locomotive and exploratory behaviour (Maier and Watkins, 1998). Both of these behaviours play a part in reward-seeking behaviour, and are known to correlate with high dopamine concentrations (Spielewoy et al., 2000).

At the biological level, cytokines known to induce sickness behaviour (IL1, IL2, IL6), also modulate the firing rate of dopamine when administered centrally (Song, 1999), (Petitto, 1997), (Ye et al., 2005).

Together both of these strands of evidence indicate that there is a functional link between immune response and dopaminergic reinforcement learning.

In addition to the effect of an immune response on the central nervous system, the central nervous system also exerts a strong influence on immune responses. Ader and Cohen (1982) famously showed that immune responses could be conditioned such that stimuli perceived in the central nervous system could come to suppress immune responses in the periphery. In their experiment conditioned stimuli (the taste of saccharine) was paired with an immunosuppressant and sheep red blood cells (SRBC). Animals later showed an aversion to saccharine, but crucially, when re-exposed to saccharine, they showed an attenuated anti-SRBC antibody response (Ader and Cohen, 1993). This attenuated response indicates that the taste of saccharine alone, a stimulus usually assumed to be perceived in the central nervous system, was now capable of suppressing the animal's immune response.

So, if the immune system modulates brain reward systems and brain reward systems modulate immune responses, then does that mean that they form part of a single circuit? Not necessarily - if one takes a narrow definition of the immune system, for example defining it only in terms of antibody production, then the causal relationship between these two systems is distant and difficult to determine.

However, it is clear that cytokines play an important role in instigating and signalling immune responses in the periphery, and that these same signalling molecules when found in the brain have a strong effect on the brain reward systems (for a review see (Dunn, 2006)).

This begs the question - How are we to interpret the action of these cytokine modu-

lators when found in the brain? Are we to treat them along with other neuromodulators as just another contingency on reward? Or do they form part of the brain's immune system? One perspective would be to interpret the actions of cytokines and inflammatory responses in the brain using the same metaphor as is used to interpret immune responses - that of defining self and non-self.

7.6.2 Self and non-self in Immunity and reward

The self/non-self metaphor was first introduced by Macfarlane Burnet, an immunologist working around the time of the second world war (Tauber, 2010). After the war there had been a sharp increase in the number of transplants, and as a result explaining the host-versus-graft became a priority in immunology reactions that were often observed. In some cases the patient's immune system would tolerate the graft from a foreign donor, but in other cases the graft would trigger an autoimmune reaction whereby the host rejected the foreign material.

At the time the immune system was seen as a reactive defence mechanism which was either passively silent, or actively attacking foreign invaders. Macfarlane Burnet proposed that tolerance and autoimmunity were two sides of the same coin and that the type of immune response that is observed depends upon whether the foreign tissue was recognised as compatible with the immune identity (ie. self or non-self). If the tissue was recognised as compatible with the self identity it would be tolerated, whilst if it was incompatible it would be rejected or destroyed.

By the late 1970s, the self/non-self metaphor introduced by Macfarlane Burnet had become the foundation for theorising in immunology, and the field had come to describe itself as the science of the self/non-self distinction (Tauber, 2010).

Since Burnet first introduced this metaphor it has become clear that the boundaries between self and non-self constituents are not straightforward, and the response of the immune system to these constituents is not always predictable. Despite these issues it is generally accepted that the self/non-self metaphor has provided fertile ground for theorising in immunology, and has undoubtedly proved practically useful, even if not true (Tauber, 2010).

While the self and non-self metaphor is most often applied at the level of antibodies, it is equally possible to apply it to the constituents of the inflammatory response generated by antibodies. Many of these constituents cross the blood-brain barrier (Quan and Banks, 2007), and in many cases they are also synthesised in neural and

CYTOKINES AND THEIR RECEPTORS KNOWN TO BE EXRESSED IN THE BRAIN

Cytokines		Cell-types			
Name	Type	Neurons	Astrocytes	Oligodendrocytes	Microglia
IL-1	Th1	C; R	C; R	C; R	C
IL-2	Th1	C; R	ND	R	R
IL-3	Th1≈Th2	C; R	C; R	R	R
IL-4	Th2	ND	R	R	R
IL-5	Th2	C; R	C; R	ND	C; R
IL-6	Th1≈Th2	C; R	C; R	ND	C; R
IL-7	ND	ND	R	R	R
IL-8	ND	R	C; R	ND	R
IL-9	Th2	ND	R	ND	ND
IL-10	Th2	C; R	ND	ND	C; R
IL-11	ND	C	ND	ND	ND
IL-12	Th1	ND	ND	ND	C
IL-13	Th2	ND	ND	ND	ND
IL-14	Th2	ND	ND	ND	ND
IL-15	Th1	C	ND	ND	C
IL-16	Th1	ND	ND	ND	ND
IL-17	Th1	ND	ND	ND	ND
IL-18	Th1	ND	ND	ND	ND
TNF-α	Th1	R	C; R	R	C; R
IFN-γ	Th1		C; R		R
TGF-β	Th2		C; R	C; R	C; R
GM-CSF	ND	R	C; R	R	R
M-CSF	ND	ND	C	R	C; R

C, cytokine; R, receptor; ND, no data; Th1≈Th2, the dominance is depending on the cell type and on the conditions; an observation that, although in a less explicit form, is possibly valid for many more cases too.

Figure 7.2: The expression of cytokines and their receptors across different cell types in the brain. Table taken from (Szelényi, 2001)

glial cells of the central nervous system (see Figure 7.2 for more details).

If the cytokines involved in defining the self and non-self in the periphery are also constitutively expressed in the central nervous system, then can we also interpret the actions of central cytokines in terms of self and non-self?

The reason for pursuing this line of argument is that concepts like reward and aversion imply the existence of a self that is rewarded and punished, and that rewarding and aversive circuits are themselves shaped and modulated by the products of inflammatory reactions (Dunn, 2006).

The interactions between inflammatory responses and neural processes are not restricted to reward - there is evidence that inflammatory cytokines play a constitutive role in memory (Jankowsky et al., 2000), (Stellwagen and Malenka, 2006), (McAfoose and Baune, 2009). Again, the blurring of these traditional boundaries between immunity and cognition suggests the need for a new model (McAfoose and Baune, 2009).

The way in which inflammatory and neural processes are interlinked suggests that we need a new metaphor to understand how and why they interact. Use of the self/non-self metaphor would allow us to interpret the function of these two systems in terms of

a larger network that serves to define the individual's identity. From this perspective, neural dynamics in the cortex can be seen as an outcome of reward learning, and in turn, reward learning can be seen as an outcome of the process of self-definition.

7.6.3 Linking the Immune and Psychological selves

The use of the self/non-self metaphor in immunology is based upon the implicit assumption that there is a biological basis to our immune identity. Likewise, in neuroscience and psychology, we frequently talk about selves — perceptual experience and cognitive control imply the existence of a self for there to be a perceiver or actor, whilst in psychiatry we often talk about disordered conceptions of self. When we do this, we usually assume that underneath all of the layers of detail there is a biological basis for the phenomena we describe.

However, it is rarely recognised that there are two concepts of self here; the immune self in the body, and the psychological self in the brain. One might wonder how these two concepts of self relate to one another — is the immune self compatible with the psychological self? If these two concepts of self interact, do the biological mechanisms of self in both cases act to support one another, do they interfere, or do they act entirely independently?

The purpose of this proposal to make use of the self/non-self metaphor in neuroscience is to suggest that *if they interact at all, they must work in tandem.*

If the biological bases of the psychological and immune selves were to be defined independently, then there would be cases where what is defined as beneficial for the immune self leads to negative outcomes for the psychological self. While this may occur in pathological cases, *this cannot occur in a psychologically and physically healthy individual*.

7.6.4 Dopamine and reward as a self(ish) circuit

We have reviewed in section 7.3 how dopamine is commonly described as a reward signal. It is most often assumed that dopamine neurons distinguish between the reward of oneself and reward of another, and so by associating dopamine with reward we are also stating that dopamine effectively distinguishes self and non-self.

In this sense we can characterise dopamine modulation as a self(ish) behavioural circuit. This would be in contrast with other systems which may motivate altruistic

actions - mirror neurons in the pre-motor cortex indicate that such concepts as "other" are already represented neurally (Rizzolatti and Craighero, 2004).

The hypothesis that dopamine neurons serve to define self does raise some interesting predictions. It is not clear if it has been experimentally verified that dopamine neurons are more responsive to the reward of self than other - evidence in the affirmative would not be surprising, but would serve to support the hypothesis.

But there may be other, more subtle ways of ascertaining whether or not dopamine neurons help define the organism's sense of self. It would be an intriguing test whether or not the cognitive sense of selfhood that is extended to objects (ie. one's own property) would be reflected in the firing of dopamine neurons. For example, we might ask: *is being rewarded with one's own property as effective at causing dopamine neurons to burst than when one is rewarded with property belonging to another?*

The proposal in this chapter is not the first time a link has been made between the firing of dopamine neurons and agency. Redgrave et al. (2008) have proposed that dopamine neurons do not signal unexpected reward, but agency — they have suggested that dopamine neurons highlight unexpected events for which the organism it*self* was responsible.

This is an interesting proposal, because it is compatible not only with the view of dopamine as signalling reward, but also fits in with dopamine's proposed role in defining a cognitive sense of self.

7.6.5 Summary

In summary, applying the self/non-self metaphor to the study of the central nervous system may offer a new perspective on processes like dopamine modulation.

Not only does the self/non-self metaphor give us a means to formulate and answer important questions about the neural basis of our sense of self, but it also makes the links between immunity and the brain more transparent.

Drawing analogies between immunity and neuroscience is valuable whether or not there is strong evidence for direct causal interactions between the two systems. An analogy between the two systems is valuable because theorists working in both fields have already developed models of learning and memory which can be applied to either domain. For very little cost we can take what has been learned in one domain and apply it to another.

Jerne (1974)'s idiotypic model of immunity can be used in neuroscience as a non-

neural model of learning and memory - a model that may become increasingly necessary as we come to recognise the ubiquity of non-neural memory systems (Kim and Linden, 2007). Quantitative examples of the workings of such a memory model can be found in Perelson and Weisbuch (1997).

Likewise, models in neuroscience may have a lot to offer immunology - reinforcement learning could be used as a model to explain how cytokines drive learning of the appropriate anti-body repertoire in the same way that dopamine drives learning of the appropriate neural response.

This would not be the first time models from immunology have made their way into neuroscience - Burnet's Clonal Selection theory and Edelman's Neural Darwinism bear some remarkable similarities.

One criticism that has been levelled at the self/non-self metaphor is that it is difficult to pin down, and this is an ongoing source of debate in immunology (Tauber, 2010). However, the difficulty we have in defining constituents of selfhood can be seen as analogous to the difficulty of defining the constituents of computation in the nervous system.

Foundational metaphors must be loosely defined if they are to be broadly applicable. Even critics of the paradigm in immunology concede that the self/non-self metaphor has proved useful (Tauber, 2000).

7.7 Dopamine: reward, wanting, or agency?

In conclusion, we have now looked at dopamine and reward in the context of two other fundamental behavioural circuits - stress and immunity. By looking at the actions of dopamine using the theoretical models of these different fields we have gained a new perspective on the role of dopamine.

When we look at the actions of dopamine *from the perspective of stress* we can see that *dopamine forms part of an anticipatory stress circuit*. Also, *dopamine can be seen as an allostatic mechanism driving the organism to engage in behaviour that takes it away from homeostasis*. The notion of dopamine as an anticipatory stress circuit, and reward as a stressor free us from the confusion over whether dopamine is involved in rewarding or aversive processing, or both.

Another view on the function of dopamine modulation comes from Berridge et al. (2009), who argue that liking and wanting can be dissociated, and that dopamine is involved in only the "wanting" component of a reward, rather than "liking" which is

mediated by opiod and endocannabinoid circuits.

Recent results support this - a study by Flagel et al. (2011) suggests that dopamine is signalling incentive salience rather than solely signalling reward prediction error. Other authors have suggested that dopamine neurons exist in two distinct populations which can encode motivational value or motivational salience (Bromberg-Martin et al., 2010).

The idea that dopamine is signalling wanting rather than liking would go some way to explaining why, when under stress, dopaminergic circuits can drive an individual to engage in behaviour they no longer gain pleasure from, a phenomenon which is common in addiction (Berridge et al., 2009). In fact it has been proposed that the priming of allostatic circuits by stress, and the consequent accumulation of allostatic load, forms the basis of addictive behaviour (Koob and Le Moal, 2001).

In addition to the context on reward offered by stress, we have also looked at dopamine function from the perspective of the self/non-self metaphor used in immunology. In this context it was suggested that dopamine can also be interpreted as signalling agency, and in the process of this, distinguishing self from non-self. This view fits well with the arguments made by Redgrave et al. (2008), who have suggested that the function of phasic dopamine is to trigger learning about changes in the environment for which the organism itself may have been responsible.

In summary the value of putting dopamine modulation and reward in context is that it allows us to see how reward-seeking behaviour can be shaped by other evolutionary critical processes going on in the body. By understanding dopamine function in the context of other systems it will be easier to explain how these systems interact on a mechanistic level. And on the behavioural level, a recognition of the context of reward-seeking behaviour will help us better understand how an individual's self-interested drive for reward relates to and is shaped by other aspects of human behaviour.

Chapter 8

Conclusion

8.1 Summary of findings

In this thesis I have taken an interesting new model in neuroscience and tried to verify it using computational models. I was unable to support or contradict the hypothesis with my simulations, but in the process of constructing and analysing the models I have learned a great deal about the difficulties of linking computational models with empirical data.

My conclusions have been

1. That it is not possible to develop cross-level models in neuroscience that can accurately relate high-level concepts to low-level empirical observations.

2. That high-level models can be useful even if they are not supported by and underlying mechanistic model.

3. That metaphors like computation, reinforcement learning, or self/non-self *must be* loosely defined if they are to be generic enough to explain many different phenomena.

I have also described other potential models and metaphors in neuroscience — such as stress, allostasis, and the self/non-self metaphor in chapter 7 — and argued that these models may be useful for understanding phenomena in which both the brain and body are involved.

8.2 How we can compare models

In the first of my conclusions above I have argued that is not possible to determine the truth of a theoretical framework in so far that it cannot be uniquely justified by empirical data. This should not be mistaken for saying that all models are as good as one another. The second question in this thesis asked whether or not neuromodulation and reinforcement was a better model for explaining behaviour when compared with neural computation. How can we judge which is a better model if neither can be verified by empirical data? In this thesis I have argued that alongside empirical arguments, there are also a priori factors that can make one model more suitable than another. A good model should:

1. Explain phenomena using variables of a similar timescale
2. Use a metaphor which is appropriate for the target domain (in section 1.3.2 I argued that this was one of the reasons computation was a poor metaphor for describing emotion and motivational states).

For these reasons I would still argue that neuromodulation and reinforcement learning is a more appropriate model for relating brain and *behaviour*, primarily because the release and effect of neuromodulation manifests itself at a timescale closer to the behaviour we are interested in. It is also more interesting from a psychological perspective because it attempts to describe our behaviour in terms of desire rather than purely mechanistic responses.

8.3 Relating brain processes to human psychology

After spending several years working on this topic I feel I have a better understanding of my motivations for pursuing this thesis. In hindsight, one of the reasons why I have been interested in neuromodulation and reinforcement learning as a model may be because it offers a description of neuroscience in which the behaviour of human beings is governed by drives and desires. I have followed this path in reaction to a model which describes humans as agents with internal representations, but no internal *motivational states*. The growing trend I have described, of researchers moving from explanations based upon neural computation to explanations based upon reinforcement learning represents a need on the part of researchers to put our desire at the centre of our behaviour. On the other hand, the metaphor of selfhood which I proposed in chapter 7

represents a need to put our identity, self-preservation, and our need assert our identity as the core of our behaviour.

Now, at the end of this thesis I believe that it is not a question of which metaphor is correct, but which metaphor a) suits the needs of the research community, or b) suits the needs of society as a whole. I would argue for the value of model pluralism, rather than a choice of one metaphor over another. Where there is a choice of models one should be free to choose the model that is most effective, or most appropriate.

8.4 Hiding the world in the brain

At the start of this thesis I set out with questions about how the brain works. My attempts to answer these questions were unsuccessful, and I have spent some time trying to understand why these models failed. With the benefit of time it now seems strange that I ever thought the models could succeed. Why is it so difficult to relate the brain and behaviour?

It may be because when we do neuroscience we make the assumption that we can relate our perception and behaviour to correlates in the brain. But by making this assumption we commit ourself to the view that all things which we can perceive or act upon (ie. all things in the world), must also have a correlate in the brain. Neuroscience must then explain how it is that the brain can contain not only all the things that exist in the world, but also all the relationships between all the things that exist in the world. When seen this way, that we are trying to hide the world in the brain, it becomes clear why our task is so difficult.

Appendix A

Chapter 3 model parameters

	a	z
Regular spiking neurons	0.02	1
Fast spiking interneurons	0.1	2

Table A.1: Parameter values for the two neuron types in the network.

Parameter	Parameter use	value
τ_c	Eligibility trace time constant	1.05
τ_d	Dopamine decay time constant	1.005

Table A.2: Parameter values for the model described in chapter 3.

Appendix B

Chapter 4 model code

```
1  // Spiking neural network with axonal conduction delays, STDP and
       dopamine
2  // Created by Eugene M. Izhikevich, Sept 23, 2005, San Diego, CA
3  // Loads values from initial3600.dat and implements secondary
       conditioning.
4  // Saves spiking data each second in file spikes.dat
5  // To plot spikes, use MATLAB code: load spikes.dat;plot(spikes(:,1)
       ,spikes(:,2),'.');
6
7  // Additional variables for analysis added by Robert Kyle, 2007,
       Edinburgh, UK
8
9  #include <iostream.h>
10 #include <math.h>
11 #include <stdio.h>
12 #include <stdlib.h>
13
14 #define getrandom(max1) ((rand()(int((max1))))) // random integer
       between 0 and max-1
15 #define T (7000)          // duration of simulation
16 int    rand_seed=0;
17
18 const int    Ne = 800;    // excitatory neurons
19 const int    Ni = 200;    // inhibitory neurons
20 const int    N  = Ne+Ni;   // total number of neurons
21 const int    M  = 100;    // the number of synapses per neuron
22 const int    D  = 1;      // maximal axonal conduction delay
23 float sm = 4.0;           // maximal synaptic strength
24 int    post[N][M];        // indeces of postsynaptic neurons
```

```
25 float s[N][M], sd[N][M];     // matrix of synaptic weights and their
       derivatives
26 short delays_length[N][D];  // distribution of delays
27 short delays[N][D][M];      // arrangement of delays
28 int   N_pre[N], I_pre[N][3*M], D_pre[N][3*M]; // presynaptic
       information
29 float *s_pre[N][3*M], *sd_pre[N][3*M];     // presynaptic weights
30 float LTP[N][1001+D], LTD[N]; // STDP functions
31 float a[N], d[N];         // neuronal dynamics parameters
32 float v[N], u[N];         // activity variables
33 int   N_firings;          // the number of fired neurons
34 const int N_firings_max=100*N;  // upper limit on the number of
       fired neurons per sec
35 int   firings[N_firings_max][2]; // indeces and timings of spikes
36
37
38 // Parameters for the secondary conditioning
39 #define VTA   (100) // neurons projecting to VAT, 0..VTA-1
40 #define US0   (100)
41 #define US1   (200) // Neurons US0..US1-1 are stimulated by the
       unconditional stimulus (US)
42 #define CSA0  (200)
43 #define CSA1  (300) // Neurons CSA0..CSA1-1 are stimulated by the
       conditional stimulus A
44 #define CSB0  (300)
45 #define CSB1  (400) // Neurons CSB0..CSB1-1 are stimulated by the
       conditional stimulus B
46 #define Tstimmin (10) // Minimal period (sec) of time between random
        trials
47 #define Tstimmax (30) // Maximal period (sec) of time between random
        trials
48 int   Tstim=10;   // The scheduled time of stimulation
49 #define CSAjitter (250) // Jitter (ms) of presentation of CSA. (
       t_CSA = t_US-1sec +CSAjit)
50 int   CSAjit=0;   // Scheduled jitter for the next trial
51 #define CSBjitter (250) // Jitter (ms) of presentation of CSB. (
       t_CSB = t_US-2sec +CSBjit)
52 int   CSBjit=0;   // Scheduled jitter for the next trial
53
54 #define StartUS (0)
55 #define StopUS  (T)
56 #define StartA  (2000)
```

```
57 #define StopA (T)
58 #define StartB  (4000)
59 #define StopB (T)
60
61 #define DAamp (0.005) // the amplitude of the reward per each spike
       of VTA neurons
62 float DA=0.0;       // concentration of extracellular DA
63 float US2US, US2VTA, US2other, US2CSa, US2CSb, US2inh;
64 float VTA2US, VTA2VTA, VTA2other, VTA2CSa, VTA2CSb, VTA2inh;
65 float other2US, other2VTA, other2other, other2CSa, other2CSb,
       other2inh;
66 float CSa2US, CSa2VTA, CSa2other, CSa2CSa, CSa2CSb, CSa2inh;
67 float CSb2US, CSb2VTA, CSb2other, CSb2CSa, CSb2CSb, CSb2inh;
68
69 int countUS2US=0,countUS2VTA=0,countUS2other=0,countUS2CSa=0,
       countUS2CSb=0, countUS2inh=0;
70 int countVTA2US=0,countVTA2VTA=0,countVTA2other=0,countVTA2CSa=0,
       countVTA2CSb=0, countVTA2inh=0;
71 int countother2US=0,countother2VTA=0,countother2other=0,
       countother2CSa=0,countother2CSb=0, countother2inh=0;
72 int countCSa2US=0,countCSa2VTA=0,countCSa2other=0,countCSa2CSa=0,
       countCSa2CSb=0, countCSa2inh=0;
73 int countCSb2US=0,countCSb2VTA=0,countCSb2other=0,countCSb2CSa=0,
       countCSb2CSb=0, countCSb2inh=0;
74
75
76
77 void initialize()
78 { int i,j,k,jj,dd, exists, r;
79   for (i=0;i<Ne;i++) a[i]=0.02;// RS type
80   for (i=Ne;i<N;i++) a[i]=0.1; // FS type
81
82   for (i=0;i<Ne;i++) d[i]=8.0; // RS type
83   for (i=Ne;i<N;i++) d[i]=2.0; // FS type
84
85   for (i=0;i<N;i++) for (j=0;j<M;j++)
86   {
87     do{
88       exists = 0;   // avoid multiple synapses
89       if (i<Ne) r = getrandom(N);
90       else      r = getrandom(Ne);// inh -> exc only
91       if (r==i) exists=1;                  // no self-synapses
```

```
92        for (k=0;k<j;k++) if (post[i][k]==r) exists = 1;  // synapse
              already exists
93      }while (exists == 1);
94      post[i][j]=r;
95    }
96    for (i=0;i<Ne;i++)  for (j=0;j<M;j++) s[i][j]=1.0;  // initial exc
        . synaptic weights
97    for (i=Ne;i<N;i++)  for (j=0;j<M;j++) s[i][j]=-1.0; // inhibitory
        synaptic weights
98      for (i=0;i<N;i++) for (j=0;j<M;j++) sd[i][j]=0.0; // synaptic
          derivatives
99      for (i=0;i<N;i++)
100   {
101     short ind=0;
102     if (i<Ne)
103     {
104       for (j=0;j<D;j++)
105       { delays_length[i][j]=M/D;  // uniform distribution of exc.
            synaptic delays
106         for (k=0;k<delays_length[i][j];k++)
107           delays[i][j][k]=ind++;
108       }
109     }
110     else
111     {
112       for (j=0;j<D;j++) delays_length[i][j]=0;
113       delays_length[i][0]=M;       // all inhibitory delays are 1 ms
114       for (k=0;k<delays_length[i][0];k++)
115           delays[i][0][k]=ind++;
116     }
117   }
118
119     for (i=0;i<N;i++)
120   {
121     N_pre[i]=0;
122     for (j=0;j<Ne;j++)
123     for (k=0;k<M;k++)
124     if (post[j][k] == i)     // find all presynaptic neurons
125     {
126       I_pre[i][N_pre[i]]=j; // add this neuron to the list
127       for (dd=0;dd<D;dd++)  // find the delay
128         for (jj=0;jj<delays_length[j][dd];jj++)
```

```
129             if (post[j][delays[j][dd][jj]]==i) D_pre[i][N_pre[i]]=dd;
130         s_pre[i][N_pre[i]]=&s[j][k];  // pointer to the synaptic
                weight
131         sd_pre[i][N_pre[i]++]=&sd[j][k];// pointer to the derivative
132       }
133     }
134
135     for (i=0;i<N;i++) for (j=0;j<1+D;j++) LTP[i][j]=0.0;
136     for (i=0;i<N;i++) LTD[i]=0.0;
137     for (i=0;i<N;i++) v[i]=-65.0;    // initial values for v
138     for (i=0;i<N;i++) u[i]=0.2*v[i];  // initial values for u
139
140     N_firings=1;     // spike timings
141     firings[0][0]=-D; // put a dummy spike at -D for simulation
          efficiency
142     firings[0][1]=0;  // index of the dummy spike
143 }
144
145 void save(char fname[30])
146 {
147   FILE *f;
148   f = fopen( fname, "wb" );
149   fwrite(&rand_seed,sizeof(int),1,f);
150   fwrite(v,sizeof(float),N,f);
151   fwrite(u,sizeof(float),N,f);
152   fwrite(s,sizeof(float),N*M,f);
153   fwrite(sd,sizeof(float),N*M,f);
154   fwrite(LTP,sizeof(float),N*(1001+D),f);
155   fwrite(LTD,sizeof(float),N,f);
156   fwrite(&N_firings,sizeof(int),1,f);
157   fwrite(firings,sizeof(int),N_firings_max*2,f);
158   fclose(f);
159 }
160 void load(char fname[30])
161 {
162   FILE *f;
163   f = fopen( fname, "rb" );
164   fread(&rand_seed,sizeof(int),1,f);
165   srand(rand_seed);
166   initialize(); // assign connections the same way as in initial3600
        .cpp  // Ah-Ha!  becuase random seed is the same, network will
        be the same
```

```
167   fread(v,sizeof(float),N,f);
168   fread(u,sizeof(float),N,f);
169   fread(s,sizeof(float),N*M,f);
170   fread(sd,sizeof(float),N*M,f);
171   fread(LTP,sizeof(float),N*(1001+D),f);
172   fread(LTD,sizeof(float),N,f);
173   fread(&N_firings,sizeof(int),1,f);
174   fread(firings,sizeof(int),N_firings_max*2,f);
175   fclose(f);
176 }
177
178
179 int main()
180 {
181   int   i, j, k, sec, t;
182   float I[N];
183   FILE  *fs;
184
185   load("initial3600.dat");          // use initial3600.cpp to generate
          the file
186
187   remove("data.dat");
188   remove("spikesUS.dat");
189   remove("spikesA.dat");
190   remove("spikesB.dat");
191
192   // US first
193     for (i=US0;i<US1;i++) for (j=0;j<M;j++)
194       if (post[i][j] <US1 && post[i][j] >=US0) countUS2US++;  // US
            target
195     for (i=US0;i<US1;i++) for (j=0;j<M;j++)
196       if (post[i][j] <VTA) countUS2VTA++;          // VTA target
197     for (i=US0;i<US1;i++) for (j=0;j<M;j++)
198       if (post[i][j] <Ne && post[i][j] >=CSB1) countUS2other++; //
            non-reward contingent target
199     for (i=US0;i<US1;i++) for (j=0;j<M;j++)
200       if (post[i][j] <CSA1 && post[i][j] >=CSA0) countUS2CSa++; //
            CSa target
201     for (i=US0;i<US1;i++) for (j=0;j<M;j++)
202       if (post[i][j] <CSB1 && post[i][j] >=CSB0) countUS2CSb++; //
            CSb target
203     for (i=US0;i<US1;i++) for (j=0;j<M;j++)
```

```
204       if (post[i][j] <N && post[i][j] >=Ne) countUS2inh++;  // inh
              target
205
206     // then VTA
207     for (i=0;i<VTA;i++) for (j=0;j<M;j++)
208       if (post[i][j] <US1 && post[i][j] >=US0) countVTA2US++; // US
              target
209     for (i=0;i<VTA;i++) for (j=0;j<M;j++)
210       if (post[i][j] <VTA) countVTA2VTA++;            // VTA target
211     for (i=0;i<VTA;i++) for (j=0;j<M;j++)
212       if (post[i][j] <Ne && post[i][j] >=CSB1) countVTA2other++;  //
              non-reward contingent target
213     for (i=0;i<VTA;i++) for (j=0;j<M;j++)
214       if (post[i][j] <CSA1 && post[i][j] >=CSA0) countVTA2CSa++;  //
              CSa target
215     for (i=0;i<VTA;i++) for (j=0;j<M;j++)
216       if (post[i][j] <CSB1 && post[i][j] >=CSB0) countVTA2CSb++;  //
              CSb target
217     for (i=0;i<VTA;i++) for (j=0;j<M;j++)
218       if (post[i][j] <N && post[i][j] >=Ne) countVTA2inh++; // inh
              target
219
220     //then other
221     for (i=CSB1;i<Ne;i++) for (j=0;j<M;j++)
222       if (post[i][j] <US1 && post[i][j] >=US0) countother2US++; //
              US target
223     for (i=CSB1;i<Ne;i++) for (j=0;j<M;j++)
224       if (post[i][j] <VTA) countother2VTA++;            // VTA target
225     for (i=CSB1;i<Ne;i++) for (j=0;j<M;j++)
226       if (post[i][j] <Ne && post[i][j] >=CSB1) countother2other++;
              // non-reward contingent target
227     for (i=CSB1;i<Ne;i++) for (j=0;j<M;j++)
228       if (post[i][j] <CSA1 && post[i][j] >=CSA0) countother2CSa++;
              // CSa target
229     for (i=CSB1;i<Ne;i++) for (j=0;j<M;j++)
230       if (post[i][j] <CSB1 && post[i][j] >=CSB0) countother2CSb++;
              // CSb target
231     for (i=CSB1;i<Ne;i++) for (j=0;j<M;j++)
232       if (post[i][j] <N && post[i][j] >=Ne) countother2inh++; // inh
              target
233
234     //then CSa
```

```
235     for (i=CSA0;i<CSA1;i++) for (j=0;j<M;j++)
236       if (post[i][j] <US1 && post[i][j] >=US0) countCSa2US++; // US
              target
237     for (i=CSA0;i<CSA1;i++) for (j=0;j<M;j++)
238       if (post[i][j] <VTA) countCSa2VTA++;            // VTA target
239     for (i=CSA0;i<CSA1;i++) for (j=0;j<M;j++)
240       if (post[i][j] <Ne && post[i][j] >=CSB1) countCSa2other++;  //
              non-reward contingent target
241     for (i=CSA0;i<CSA1;i++) for (j=0;j<M;j++)
242       if (post[i][j] <CSA1 && post[i][j] >=CSA0) countCSa2CSa++;  //
              CSa target
243     for (i=CSA0;i<CSA1;i++) for (j=0;j<M;j++)
244       if (post[i][j] <CSB1 && post[i][j] >=CSB0) countCSa2CSb++;  //
              CSb target
245     for (i=CSA0;i<CSA1;i++) for (j=0;j<M;j++)
246       if (post[i][j] <N && post[i][j] >=Ne) countCSa2inh++; // inh
              target
247
248     // CSb
249     for (i=CSB0;i<CSB1;i++) for (j=0;j<M;j++)
250       if (post[i][j] <US1 && post[i][j] >=US0) countCSb2US++; // US
              target
251     for (i=CSB0;i<CSB1;i++) for (j=0;j<M;j++)
252       if (post[i][j] <VTA) countCSb2VTA++;            // VTA target
253     for (i=CSB0;i<CSB1;i++) for (j=0;j<M;j++)
254       if (post[i][j] <Ne && post[i][j] >=CSB1) countCSb2other++;  //
              non-reward contingent target
255     for (i=CSB0;i<CSB1;i++) for (j=0;j<M;j++)
256       if (post[i][j] <CSA1 && post[i][j] >=CSA0) countCSb2CSa++;  //
              CSa target
257     for (i=CSB0;i<CSB1;i++) for (j=0;j<M;j++)
258       if (post[i][j] <CSB1 && post[i][j] >=CSB0) countCSb2CSb++;  //
              CSb target
259     for (i=CSB0;i<CSB1;i++) for (j=0;j<M;j++)
260       if (post[i][j] <N && post[i][j] >=Ne) countCSb2inh++; // inh
              target
261
262
263     fs= fopen("data.dat","a");
264     fprintf(fs," d fprintf(fs," fprintf(fs," fprintf(fs," fclose(fs;
265
```

```
266  for (i=US0;i<US1;i++) for (j=0;j<M;j++) if (post[i][j]<VTA) s[i][j
         ]=0.75*sm; //unconditioned reward    // US->VTA set to 3, not 4
267
268  for (sec=0; sec<T; sec++)
269  {
270    for (t=0;t<1000;t++)          // simulation of 1 sec
271    {
272      for (i=0;i<N;i++) I[i] = 13.0*(getrandom(1000)/1000.0-0.5); //
             reset the input
273
274      // The following code does: US, CSa+US, and CSb+CSa+US.
275
276      if ((sec==Tstim-2) & (t==500+CSBjit) & (sec>StartB) & (sec<=
             StopB)) // stimulation of CSB
277      {
278        for (i=CSB0;i<CSB1;i++) I[i]=20.0;
279      }
280      if ((sec==Tstim-1) & (t==500+CSAjit) & (sec>StartA) & (sec<=
             StopA)) // stimulation of CSA
281      {
282          for (i=CSA0;i<CSA1;i++) I[i]=20.0;
283      }
284      if ((sec==Tstim) & (t==500) & (sec>StartUS) & (sec<=StopUS))
             // stimulation of US
285      {
286        for (i=US0;i<US1;i++) I[i]=20.0;       // superthreshold
             current stimulation
287      }
288
289      for (i=0;i<N;i++)
290      if (v[i]>=30)          // did it fire?
291      {
292
293        if (i<VTA) DA+=DAamp;    // Add 0.005 per VTA spike
294
295        v[i] = -65.0;          // voltage reset
296        u[i]+=d[i];            // recovery variable reset
297        LTP[i][t+D]= 0.1;
298        LTD[i]=0.15;
299        for (j=0;j<N_pre[i];j++) *sd_pre[i][j]+=LTP[I_pre[i][j]][t+D
             -D_pre[i][j]-1];// this spike was after pre-synaptic
             spikes
```

```
300           firings[N_firings  ][0]=t;
301           firings[N_firings++][1]=i;
302           if (N_firings == N_firings_max) {cout << "Two many spikes at
                  t=" << t << " (ignoring all)";N_firings=1;}
303         }
304         k=N_firings;
305         while (t-firings[--k][0] <D)
306         {
307           for (j=0; j< delays_length[firings[k][1]][t-firings[k][0]];
                j++)
308           {
309             i=post[firings[k][1]][delays[firings[k][1]][t-firings[k
                    ][0]][j]];
310             I[i]+=s[firings[k][1]][delays[firings[k][1]][t-firings[k
                    ][0]][j]];
311             if (firings[k][1] <Ne) // this spike is before
                    postsynaptic spikes
312               sd[firings[k][1]][delays[firings[k][1]][t-firings[k
                      ][0]][j]]-=LTD[i];
313           }
314         }
315         for (i=0;i<N;i++)
316         {
317           v[i]+=0.5*((0.04*v[i]+5)*v[i]+140-u[i]+I[i]); // for
                numerical stability
318           v[i]+=0.5*((0.04*v[i]+5)*v[i]+140-u[i]+I[i]); // time step
                is 0.5 ms
319           u[i]+=a[i]*(0.2*v[i]-u[i]);
320           LTP[i][t+D+1]=0.95*LTP[i][t+D];
321           LTD[i]*=0.95;
322         }
323
324         DA*=0.995;
325
326         if (t10==0      // do every 10th ms for simulation efficiency
327         {
328           for (i=0;i<Ne;i++)  // modify only exc connections
329           for (j=0;j<M;j++)
330           {
331             s[i][j]+=sd[i][j]*(0.00+DA);     // No ad-hoc tonic DA this
                  time
332             if (s[i][j]>sm) s[i][j]=sm;
```

```
333             if (s[i][j]<0) s[i][j]=0.0;
334           }
335           for (i=0;i<Ne;i++)
336           for (j=0;j<M;j++)
337           sd[i][j]*=0.99;
338         }
339       }
340
341       int N_fir_exc=0, N_fir_inh=0;
342       for (i=1;i<N_firings;i++)
343         if (firings[i][1] <Ne)  N_fir_exc++; else N_fir_inh++;
344       fprintf(stderr,"sec=d, exc.frate=US2US=0.0; US2VTA=0.0; US2other=0.0;
              US2CSa=0.0; US2CSb=0.0; US2inh=0.0;VTA2US=0.0; VTA2VTA=0.0; VTA2other=0.0;
              VTA2CSa=0.0; VTA2CSb=0.0; VTA2inh=0.0;other2US=0.0; other2VTA=0.0;
              other2other=0.0; other2CSa=0.0; other2CSb=0.0; other2inh=0.0;CSa2US=0.0;
              CSa2VTA=0.0; CSa2other=0.0; CSa2CSa=0.0; CSa2CSb=0.0; CSa2inh=0.0;CSb2US=0.0;
              CSb2VTA=0.0; CSb2other=0.0; CSb2CSa=0.0; CSb2CSb=0.0; CSb2inh=0.0;// US firstfor
              (i=US0;i<US1;i++ for (j=0;j<M;j++)
345         if (post[i][j] <US1 && post[i][j] >=US0) US2US+=s[i][j];  //
              US target
346       for (i=US0;i<US1;i++) for (j=0;j<M;j++)
347         if (post[i][j] <VTA) US2VTA+=s[i][j];          // VTA target
348       for (i=US0;i<US1;i++) for (j=0;j<M;j++)
349         if (post[i][j] <Ne && post[i][j] >=CSB1) US2other+=s[i][j]; //
              non-reward contingent target
350       for (i=US0;i<US1;i++) for (j=0;j<M;j++)
351         if (post[i][j] <CSA1 && post[i][j] >=CSA0) US2CSa+=s[i][j]; //
              CSa target
352       for (i=US0;i<US1;i++) for (j=0;j<M;j++)
353         if (post[i][j] <CSB1 && post[i][j] >=CSB0) US2CSb+=s[i][j]; //
              CSb target
354       for (i=US0;i<US1;i++) for (j=0;j<M;j++)
355         if (post[i][j] <N && post[i][j] >=Ne) US2inh+=s[i][j];  // inh
              target
356
357       // then VTA
358       for (i=0;i<VTA;i++) for (j=0;j<M;j++)
359         if (post[i][j] <US1 && post[i][j] >=US0) VTA2US+=s[i][j]; //
              US target
360       for (i=0;i<VTA;i++) for (j=0;j<M;j++)
361         if (post[i][j] <VTA) VTA2VTA+=s[i][j];          // VTA target
362       for (i=0;i<VTA;i++) for (j=0;j<M;j++)
363         if (post[i][j] <Ne && post[i][j] >=CSB1) VTA2other+=s[i][j];
              // non-reward contingent target
364       for (i=0;i<VTA;i++) for (j=0;j<M;j++)
```

```
365       if (post[i][j] <CSA1 && post[i][j] >=CSA0) VTA2CSa+=s[i][j];
              // CSa target
366     for (i=0;i<VTA;i++) for (j=0;j<M;j++)
367       if (post[i][j] <CSB1 && post[i][j] >=CSB0) VTA2CSb+=s[i][j];
              // CSb target
368     for (i=0;i<VTA;i++) for (j=0;j<M;j++)
369       if (post[i][j] <N && post[i][j] >=Ne) VTA2inh+=s[i][j]; // inh
              target
370
371     //then other
372     for (i=CSB1;i<Ne;i++) for (j=0;j<M;j++)
373       if (post[i][j] <US1 && post[i][j] >=US0) other2US+=s[i][j]; //
              US target
374     for (i=CSB1;i<Ne;i++) for (j=0;j<M;j++)
375       if (post[i][j] <VTA) other2VTA+=s[i][j];            // VTA
              target
376     for (i=CSB1;i<Ne;i++) for (j=0;j<M;j++)
377       if (post[i][j] <Ne && post[i][j] >=CSB1) other2other+=s[i][j];
              // non-reward contingent target
378     for (i=CSB1;i<Ne;i++) for (j=0;j<M;j++)
379       if (post[i][j] <CSA1 && post[i][j] >=CSA0) other2CSa+=s[i][j];
              // CSa target
380     for (i=CSB1;i<Ne;i++) for (j=0;j<M;j++)
381       if (post[i][j] <CSB1 && post[i][j] >=CSB0) other2CSb+=s[i][j];
              // CSb target
382     for (i=CSB1;i<Ne;i++) for (j=0;j<M;j++)
383       if (post[i][j] <N && post[i][j] >=Ne) other2inh+=s[i][j]; //
              inh target
384
385     //then CSa
386     for (i=CSA0;i<CSA1;i++) for (j=0;j<M;j++)
387       if (post[i][j] <US1 && post[i][j] >=US0) CSa2US+=s[i][j]; //
              US target
388     for (i=CSA0;i<CSA1;i++) for (j=0;j<M;j++)
389       if (post[i][j] <VTA) CSa2VTA+=s[i][j];            // VTA target
390     for (i=CSA0;i<CSA1;i++) for (j=0;j<M;j++)
391       if (post[i][j] <Ne && post[i][j] >=CSB1) CSa2other+=s[i][j];
              // non-reward contingent target
392     for (i=CSA0;i<CSA1;i++) for (j=0;j<M;j++)
393       if (post[i][j] <CSA1 && post[i][j] >=CSA0) CSa2CSa+=s[i][j];
              // CSa target
394     for (i=CSA0;i<CSA1;i++) for (j=0;j<M;j++)
```

```
395       if (post[i][j] <CSB1 && post[i][j] >=CSB0) CSa2CSb+=s[i][j];
              // CSb target
396     for (i=CSA0;i<CSA1;i++) for (j=0;j<M;j++)
397       if (post[i][j] <N && post[i][j] >=Ne) CSa2inh+=s[i][j]; // inh
              target
398
399     // CSb
400     for (i=CSB0;i<CSB1;i++) for (j=0;j<M;j++)
401       if (post[i][j] <US1 && post[i][j] >=US0) CSb2US+=s[i][j]; //
          US target
402     for (i=CSB0;i<CSB1;i++) for (j=0;j<M;j++)
403       if (post[i][j] <VTA) CSb2VTA+=s[i][j];            // VTA target
404     for (i=CSB0;i<CSB1;i++) for (j=0;j<M;j++)
405       if (post[i][j] <Ne && post[i][j] >=CSB1) CSb2other+=s[i][j];
          // non-reward contingent target
406     for (i=CSB0;i<CSB1;i++) for (j=0;j<M;j++)
407       if (post[i][j] <CSA1 && post[i][j] >=CSA0) CSb2CSa+=s[i][j];
          // CSa target
408     for (i=CSB0;i<CSB1;i++) for (j=0;j<M;j++)
409       if (post[i][j] <CSB1 && post[i][j] >=CSB0) CSb2CSb+=s[i][j];
          // CSb target
410     for (i=CSB0;i<CSB1;i++) for (j=0;j<M;j++)
411       if (post[i][j] <N && post[i][j] >=Ne) CSb2inh+=s[i][j]; // inh
              target
412
413     fs= fopen("data.dat","a");
414     fprintf(fs," ffprintf(fs," fprintf(fs," fprintf(fs," fclose(fs;
415
416     fs = fopen("spikes.dat","w");
417     for (i=1;i<N_firings;i++)
418       if (firings[i][0] >=0)
419         fprintf(fs, "dfclose(fs;
420
421     if (sec==Tstim-2) // CSB
422     {
423       fs = fopen("spikesB.dat","a");
424       for (i=1;i<N_firings;i++)
425         if (abs(firings[i][0]-(500+CSBjit))<= 500-CSBjitter)  // If
              spike occurred 250ms before or after CSB ?
426           fprintf(fs, "dfclose(fs;
427     }
428     if (sec==Tstim-1) // CSA
```

```
429     {
430       fs = fopen("spikesA.dat","a");
431       for (i=1;i<N_firings;i++)
432         if (abs(firings[i][0]-(500+CSAjit))<= 500-CSAjitter)
433           fprintf(fs, "d fclose(fs;
434     }
435     if (sec==Tstim) // US
436     {
437       fs = fopen("spikesUS.dat","a");
438       for (i=1;i<N_firings;i++)
439         if (firings[i][0]>=0)
440           fprintf(fs, "d fclose(fs;
441       Tstim+=Tstimmin+getrandom(Tstimmax-Tstimmin); // Set time for
            next stimulation
442       CSAjit=getrandom(2*CSAjitter)-CSAjitter;
443       CSBjit=getrandom(2*CSBjitter)-CSBjitter;
444     }
445
446
447     for (i=0;i<N;i++)    // prepare for the next sec
448       for (j=0;j<D+1;j++)
449       LTP[i][j]=LTP[i][1000+j];
450     k=N_firings-1;
451     while (1000-firings[k][0]<D) k--; // Get last D timesteps to the
          start of the matrix, and count off the number of firings = k
452     for (i=1;i<N_firings-k;i++)
453     {
454       firings[i][0]=firings[k+i][0]-1000;
455       firings[i][1]=firings[k+i][1];
456     }
457     N_firings = N_firings-k;       // Number of firings = Number of
          firings in final timestep
458   }
459
460   fs=fopen("synapses.dat","w");
461   for (i=0;i<N;i++)
462     for (j=0;j<M;j++)
463     fprintf(fs,"f,",s[i][j];
464   fprintf(fs,"\n");
465   fclose(fs);
466
467   fs=fopen("connectivity.dat","w");
```

```
468     for (i=0;i<N;i++)
469       for (j=0;j<M;j++)
470       fprintf(fs,"f,",post[i][j];
471     fprintf(fs,"\n");
472     fclose(fs);
473
474     save("schultz_end.dat");
475     return 0;
476 }
```

Appendix C

Chapter 4 model parameters

Parameter	Parameter use	value
g_L	Leak channel maximum conductance	8.06mS
E_L	Leak channel reversal potential	-80mV
g_{Na}	Sodium-like channel maximum conductance	23mS
E_{Na}	Sodium channel reversal potential	60mV
g_K	Potassium-like channel maximum conductance	13mS
E_K	Potassium channel reversal potential	-90mV
g_M	M channel maximum conductance	19mS
E_M	M channel reversal potential	-90mV
$V_{NaInhalfmax}$	half the maximal conductance of the Sodium-like channel	-54.1
k_{NaIn}	rate constant for voltage-gating of the Sodium-like channel	-1
$V_{Khalfmax}$	half the maximal conductance of the Potassium-like channel	-25
k_K	rate constant for voltage-gating of the Potassium-like channel	5
$V_{Mhalfmax}$	half the maximal conductance of the M channel	-20
k_M	rate constant for voltage-gating of the M channel	15
τ_{NaIn}	time constant for gating variable on the sodium-like channel	500
τ_K	time constant for gating variable on the potassium-like channel	1
τ_M	time constant for gating variable on the M channel	200

Table C.1: Parameter values for the neuron model described in chapter 4.

Parameter	Parameter use	value
$\tau_{1,AMPA}$	AMPA channel opening time constant	1
$\tau_{2,AMPA}$	AMPA channel closing time constant	5
$\tau_{1,NMDA}$	NMDA channel opening time constant	2
$\tau_{2,NMDA}$	NMDA channel closing time constant	80
$g_{AMPA_{max}}$	AMPA channel maximum conductance	0.1mS
$g_{NMDA_{max}}$	NMDA channel maximum conductance	0.01mS
g_{Ca}	Calcium channel maximum conductance	1

Table C.2: Parameter values for the synapses described in chapter 4.

Parameter	Parameter use	value
α_1	Parameter specifying the plasticity curve Ω	0.35
β_1	Parameter specifying the plasticity curve Ω	80
α_2	Parameter specifying the plasticity curve Ω	0.55
β_2	Parameter specifying the plasticity curve Ω	80
P_1	Parameter defining the rate of plasticity change η	0.1
P_2	Parameter defining the rate of plasticity change η	0.00001
P_3	Parameter defining the rate of plasticity change η	3
P_4	Parameter defining the rate of plasticity change η	1

Table C.3: Parameter values for the plasticity model described in chapter 4.

Appendix D

Parameters used in preliminary dopamine diffusion and re-uptake model

Each dopamine terminal shown in Figure 2.6 acts as a point source of dopamine. The resulting dopamine concentration is the sum of the concentration from each point source. The concentration from a point source at distance r at time t after release is given by the following formula taken from (Cragg and Rice, 2004):

$$C(r,t) = \frac{C_f}{\alpha(4D^*t\pi)^{\frac{3}{2}}}\exp(\frac{-r^2}{4D^*t})\exp(-k't) \quad \text{(D.1)}$$

Where $C(r,t)$ is the extra cellular dopamine concentration following one spike as a function of distance r, and time t after release. Release is assumed to be instantaneous from a vesicle with a fill concentration of C_f. Diffusion of dopamine molecules is governed by the local extracellular volume fraction α, and the tortuosity of the extracellular media λ. Tortuosity decreases the diffusion coefficient to D^* ($D^* = D\lambda^2$). The re-uptake of dopamine via DATs and oxidisation via MAOs is incorporated by the uptake constant k'.

Parameter	Parameter use	value
C_f	Fill concentration of one vesicle	1.6274e-20M
α	striatal extracellular volume fraction	0.21
λ	Tortuosity of extracellular medium	$1.54 cm^{-1}$
D	Diffusion coefficient	7.63e-6s^{-1}
k'	Dopamine re-uptake time constant	20
$f_{pacemaking}$	firing rate of pacemaking dopamine neurons	5Hz
f_{burst}	firing rate of dopamine neurons during phasic spike	80Hz
burst fraction	fraction of dopamine neurons that fire during a phasic spike	0.8
$DA_{terminals}$	dopamine terminal density in striatum	$0.104 \mu m^{-3}$

Table D.1: Parameter values for the preliminary dopamine diffusion and re-uptake model.

Bibliography

Abbott, L. F. and Blum, K. I. (1996). Functional significance of Long-Term potentiation for sequence learning and prediction. *Cerebral Cortex*, 6(3):406–416.

Ader, R. and Cohen, N. (1982). Behaviorally conditioned immunosuppression and murine systemic lupus erythematosus. *Science*, 215(4539):1534–1536.

Ader, R. and Cohen, N. (1993). Psychoneuroimmunology: conditioning and stress. *Annu Rev Psychol*, 44:53–85.

Albus, J. S. (1971). A theory of cerebellar function. *Mathematical Biosciences*, 10(1-2):25–61.

Alle, H. and Geiger, J. R. (2006). Combined analog and action potential coding in hippocampal mossy fibers. *Science (New York, N.Y.)*, 311(5765):1290–1293.

Araque, A., Parpura, V., Sanzgiri, R. P., and Haydon, P. G. (1999). Tripartite synapses: glia, the unacknowledged partner. *Trends in neurosciences*, 22(5):208–215.

Arbuthnott, G. W. W. and Wickens, J. (2007). Space, time and dopamine. *Trends Neurosci*, 30(2):62–69.

Ardid, S., Wang, X.-J. J., and Compte, A. (2007). An integrated microcircuit model of attentional processing in the neocortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(32):8486–8495.

Aston-Jones, G. and Cohen, J. D. (2005). Adaptive gain and the role of the locus coeruleus-norepinephrine system in optimal performance. *J Comp Neurol*, 493(1):99–110.

Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite, R. E. P., Filho, W. J., Lent, R., and Herculano-Houzel, S. (2009). Equal

numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J. Comp. Neurol.*, 513(5):532–541.

Baldwin, A. E., Sadeghian, K., and Kelley, A. E. (2002). Appetitive instrumental learning requires coincident activation of NMDA and dopamine d1 receptors within the medial prefrontal cortex. *The Journal of Neuroscience*, 22(3):1063–1071.

Bar-Gad, I., Morris, G., and Bergman, H. (2003). Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Progress in Neurobiology*, 71(6):439–473.

Barroso-Chinea, P., Cruz-Muros, I., Aymerich, M. S., Rodríguez-Díaz, M., Afonso-Oramas, D., Lanciego, J. L., and González-Hernández, T. (2005). Striatal expression of GDNF and differential vulnerability of midbrain dopaminergic cells. *European Journal of Neuroscience*, 21(7):1815–1827.

Bekar, L. K., He, W., and Nedergaard, M. (2008). Locus coeruleus alpha-adrenergic-mediated activation of cortical astrocytes in vivo. *Cereb. Cortex*, pages bhn040+.

Berns, G. S. and Sejnowski, T. J. (1998). A computational model of how the basal ganglia produce sequences. *Journal of Cognitive Neuroscience*, 10(1):108–121.

Berridge, K. C., Robinson, T. E., and Aldridge, J. W. (2009). Dissecting components of reward: 'liking', 'wanting', and learning. *Current Opinion in Pharmacology*, 9(1):65–73.

Bi, G. Q. and Poo, M. M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 18(24):10464–10472.

Björklund, A. and Dunnett, S. B. (2007a). Dopamine neuron systems in the brain: an update. *Trends in Neurosciences*, 30(5):194–202.

Björklund, A. and Dunnett, S. B. (2007b). Fifty years of dopamine research. *Trends in Neurosciences*, 30(5):185–187.

Blond, O. (2002). Long-term potentiation in rat prefrontal slices facilitated by phased application of dopamine. *European Journal of Pharmacology*, 438(1-2):115–116.

Bluthe, R., Dantzer, R., and Kelley, K. (1992). Effects of interleukin-1 receptor antagonist on the behavioral effects of lipopolysaccharide in rat. *Brain Research*, 573(2):318–320.

Breland, K. and Breland, M. (1961). The misbehavior of organisms. *American Psychologist*, 16(11):681–684.

Bromberg-Martin, E. S., Matsumoto, M., and Hikosaka, O. (2010). Dopamine in motivational control: Rewarding, aversive, and alerting. *Neuron*, 68(5):815–834.

Brown, J., Bullock, D., and Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *The Journal of Neuroscience*, 19(23):10502–10511.

Buckholtz, J. W., Treadway, M. T., Cowan, R. L., Woodward, N. D., Benning, S. D., Li, R., Ansari, M. S., Baldwin, R. M., Schwartzman, A. N., Shelby, E. S., Smith, C. E., Cole, D., Kessler, R. M., and Zald, D. H. (2010). Mesolimbic dopamine reward system hypersensitivity in individuals with psychopathic traits. *Nature neuroscience*, 13(4):419–421.

Burgess, N., Barry, C., and O'Keefe, J. (2007). An oscillatory interference model of grid cell firing. *Hippocampus*, 17(9):801–812.

Calabresi, P., Picconi, B., Tozzi, A., and Di Filippo, M. (2007). Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends Neurosci*.

Carlsson, A., Lindqvist, M., and Magnusson, T. (1957). 3,4-Dihydroxyphenylalanine and 5-Hydroxytryptophan as reserpine antagonists. *Nature*, 180(4596):1200.

Celada, P., Paladini, C. A., and Tepper, J. M. (1999). GABAergic control of rat substantia nigra dopaminergic neurons: role of globus pallidus and substantia nigra pars reticulata. *Neuroscience*, 89(3):813–825.

Chomsky, N. (1959). Review of verbal behavior. *Language*, 35(1).

Compte, A., Brunel, N., Goldman-Rakic, P. S., and Wang, X.-J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, 10(9):910–923.

Contreras-Vidal, J. L. and Schultz, W. (1999). A predictive reinforcement model of dopamine neurons for learning approach behavior. *Journal of Computational Neuroscience*, 6(3):191–214.

Cragg, S. J. and Rice, M. E. (2004). DAncing past the DAT at a DA synapse. *Trends in Neurosciences*, 27(5):270–277.

Dahan, L., Astier, B., Vautrelle, N., Urbain, N., Kocsis, B., and Chouvet, G. (2006). Prominent burst firing of dopaminergic neurons in the ventral tegmental area during paradoxical sleep. *Neuropsychopharmacology*, 32(6):1232–1241.

Dawkins, R. (1990). *The Selfish Gene*. Oxford University Press.

Dayan, P. and Abbott, L. F. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press, 1st edition.

Dayan, P. and Huys, Q. J. (2008). Serotonin, inhibition, and negative mood. *PLoS computational biology*, 4(2):e4+.

Devoto, P., Flore, G., Saba, P., Fa, M., and Gessa, G. (2005). Co-release of noradrenaline and dopamine in the cerebral cortex elicited by single train and repeated train stimulation of the locus coeruleus. *BMC Neuroscience*, 6(1).

Di Matteo, V., De Blasi, A., Di Giulio, C., and Esposito, E. (2001). Role of 5-HT2C receptors in the control of central dopamine function. *Trends in Pharmacological Sciences*, 22(5):229–232.

Diana, M. and Tepper, J. (2002). *Electrophysiological Pharmacology Of Mesencephalic Dopaminergic Neurons*, chapter 13. Springer.

Doya, K. (2002). Metalearning and neuromodulation. *Neural networks : the official journal of the International Neural Network Society*, 15(4-6):495–506.

Dunn, A. (2006). Effects of cytokines and infections on brain neurochemistry. *Clinical Neuroscience Research*, 6(1-2):52–68.

Durstewitz, D., Seamans, J. K., and Sejnowski, T. J. (2000a). Dopamine-mediated stabilization of delay-period activity in a network model of prefrontal cortex. *Journal of neurophysiology*, 83(3):1733–1750.

Durstewitz, D., Seamans, J. K., and Sejnowski, T. J. (2000b). Neurocomputational models of working memory. *Nature neuroscience*, 3 Suppl:1184–1191.

Edelman, G. (1993). Neural darwinism: Selection and reentrant signaling in higher brain function. *Neuron*, 10(2):115–125.

Feigenbaum, M. J. (1978). Quantitative universality for a class of nonlinear transformations. *Journal of Statistical Physics*, 19(1):25–52.

Fernandez, E., Schiappa, R., Girault, J.-A. A., and Le Novère, N. (2006). DARPP-32 is a robust integrator of dopamine and glutamate signals. *PLoS computational biology*, 2(12).

Flagel, S. B., Clark, J. J., Robinson, T. E., Mayo, L., Czuj, A., Willuhn, I., Akers, C. A., Clinton, S. M., Phillips, P. E. M., and Akil, H. (2011). A selective role for dopamine in stimulus-reward learning. *Nature*, 469(7328):53–57.

Floresco, S. B., West, A. R., Ash, B., Moore, H., and Grace, A. A. (2003). Afferent modulation of dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. *Nat Neurosci*, 6(9):968–973.

Florian, R. V. (2007). Reinforcement learning through modulation of Spike-Timing-dependent synaptic plasticity. *Neural Computation*, 19(6):1468–1502.

Frank, M. J. and Surmeier, D. J. (2009). Do substantia nigra dopaminergic neurons differentiate between reward and punishment? *Journal of Molecular Cell Biology*, 1(1):15–16.

Franks, K. (2001). An MCell model of calcium dynamics and frequency-dependence of calmodulin activation in dendritic spines. *Neurocomputing*, 38-40(1-4):9–16.

Freud, S. (2005). *The Unconscious (Penguin Modern Classics Translated Texts)*. Penguin Books Ltd.

Fuster, J. M. and Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science (New York, N.Y.)*, 173(997):652–654.

Gariano, R. F. and Groves, P. M. (1988). Burst firing induced in midbrain dopamine neurons by stimulation of the medial prefrontal and anterior cingulate cortices. *Brain Res*, 462(1):194–198.

Geisler, S. and Zahm, D. S. (2005). Afferents of the ventral tegmental area in the rat-anatomical substratum for integrative functions. *The Journal of Comparative Neurology*, 490(3):270–294.

Giaume, C., Kirchhoff, F., Matute, C., Reichenbach, A., and Verkhratsky, A. (2007). Glia: the fulcrum of brain diseases. *Cell Death & Differentiation*, 14(7):1324–1335.

Goldman, M. S., Golowasch, J., Marder, E., and Abbott, L. F. (2001). Global structure, robustness, and modulation of neuronal models. *J. Neurosci.*, 21(14):5229–5238.

Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron*, 14(3):477–485.

Gonon, F. (1997). Prolonged and extrasynaptic excitatory action of dopamine mediated by d1 receptors in the rat striatum in vivo. *J. Neurosci.*, 17(15):5972–5978.

Gorelova, N., Seamans, J. K., and Yang, C. R. (2002). Mechanisms of dopamine activation of fast-spiking interneurons that exert inhibition in rat prefrontal cortex. *J Neurophysiol*, 88(6):3150–3166.

Grace, A. A., Floresco, S. B., Goto, Y., and Lodge, D. J. (2007). Regulation of firing of dopaminergic neurons and control of goal-directed behaviors. *Trends in neurosciences*, 30(5):220–227.

Greengard, P., Allen, P. B., and Nairn, A. C. (1999). Beyond the dopamine receptor: the DARPP-32/protein phosphatase-1 cascade. *Neuron*, 23(3):435–447.

Gruber, A. J., Dayan, P., Gutkin, B. S., and Solla, S. A. (2006). Dopamine modulation in the basal ganglia locks the gate to working memory. *J Comput Neurosci*, 20(2):153–166.

Gruber, A. J., Solla, S. A., Surmeier, D. J., and Houk, J. C. (2003). Modulation of striatal single units by expected reward: a spiny neuron model displaying dopamine-induced bistability. *J Neurophysiol*, 90(2):1095–1114.

Gurden, H., Takita, M., and Jay, T. M. (2000). Essential role of d1 but not d2 receptors in the NMDA Receptor-Dependent Long-Term potentiation at Hippocampal-Prefrontal cortex synapses in vivo. *The Journal of Neuroscience*, 20(22):RC106.

Haber, S. N., Kim, K. S., Mailly, P., and Calzavara, R. (2006). Reward-related cortical inputs define a large striatal region in primates that interface with associative

cortical connections, providing a substrate for incentive-based learning. *J Neurosci*, 26(32):8368–8376.

Hacking, I. (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press.

Hanisch, U.-K. (2002). Microglia as a source and target of cytokines. *Glia*, 40(2):140–155.

Harris, G. C., Wimmer, M., and Aston-Jones, G. (2005). A role for lateral hypothalamic orexin neurons in reward seeking. *Nature*, 437(7058):556–559.

Hebb (1949). *The organization of behavior : a neuropsychological theory / D.O. Hebb*. New York : Wiley.

Herkenham, M. (1987). Mismatches between neurotransmitter and receptor localizations in brain: observations and implications. *Neuroscience*, 23(1):1–38.

Herman, J. P., Figueiredo, H., Mueller, N. K., Ulrich-Lai, Y., Ostrander, M. M., Choi, D. C., and Cullinan, W. E. (2003). Central mechanisms of stress integration: hierarchical circuitry controlling hypothalamo-pituitary-adrenocortical responsiveness. *Frontiers in Neuroendocrinology*, 24(3):151–180.

Hollerman, J. R., Tremblay, L., and Schultz, W. (1998). Influence of reward expectation on Behavior-Related neuronal activity in primate striatum. *Journal of Neurophysiology*, 80(2):947–963.

Houk, J. C., Adams, J. L., and Barto, A. G. (1994). *A model of how the basal ganglia generate and use neural signals that predict reinforcement*. The MIT Press, 1 edition.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160:106–154.

Hull, C. L. (1943). *Principles of behavior, an introduction to behavior theory / by Clark L. Hull*. D. Appleton-Century Company, Incorporated, New York, London :.

Hurd, Y. L., Suzuki, M., and Sedvall, G. C. (2001). D1 and d2 dopamine receptor mRNA expression in whole hemisphere sections of the human brain. *J Chem Neuroanat*, 22(1-2):127–137.

Ito, M. (1984). *Cerebellum and Neural Control*. Raven Pr.

Izhikevich, E. M. (2006). *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting (Computational Neuroscience)*. The MIT Press, 1 edition.

Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cerebral Cortex*, 17(10):2443–2452.

Jacobs, B. L. and Fornal, C. A. (1995). Activation of 5-HT neuronal activity during motor behavior. *Seminars in Neuroscience*, 7(6):401–408.

Jankowsky, J. L., Derrick, B. E., and Patterson, P. H. (2000). Cytokine responses to LTP induction in the rat hippocampus: A comparison of in vitro and in vivo techniques. *Learn. Mem.*, 7(6):400–412.

Jerne, N. K. (1974). Towards a network theory of the immune system. *Annales d'immunologie*, 125C(1-2):373–389.

Johnson, S. W. and North, R. A. (1992). Two types of neurone in the rat ventral tegmental area and their synaptic inputs. *The Journal of physiology*, 450:455–468.

Kalivas, P. (1993). Neurotransmitter regulation of dopamine neurons in the ventral tegmental area. *Brain Research Reviews*, 18(1):75–113.

Kampa, B., Letzkus, J., and Stuart, G. (2007). Dendritic mechanisms controlling spike-timing-dependent synaptic plasticity. *Trends in Neurosciences*, 30(9):456–463.

Kandel, E., Schwartz, J., and Jessell, T. (2000). *Principles of Neural Science*. McGraw-Hill Medical, 4 edition.

Kandel, E. R. and Tauc, L. (1965). Mechanism of heterosynaptic facilitation in the giant cell of the abdominal ganglion of aplysia depilans. *The Journal of physiology*, 181(1):28–47.

Katz, P. S. (1999). *Beyond Neurotransmission: Neuromodulation and its Importance for Information Processing: Neuromodulation and Its Importance for Information Processing*. OUP Oxford, 1st edition.

Kim, S. J. and Linden, D. J. (2007). Ubiquitous plasticity and memory storage. *Neuron*, 56(4):582–592.

Kobayashi, Y. and Okada, K.-I. (2007). Reward prediction error computation in the pedunculopontine tegmental nucleus neurons. *Annals of the New York Academy of Sciences*, 1104(1):310–323.

Komura, Y., Tamura, R., Uwano, T., Nishijo, H., Kaga, K., and Ono, T. (2001). Retrospective and prospective coding for predicted reward in the sensory thalamus. *Nature*, 412(6846):546–549.

Koob, G. F. and Le Moal, M. (2001). Drug addiction, dysregulation of reward, and allostasis. *Neuropsychopharmacology*, 24(2):97–129.

Korotkova, T. M., Sergeeva, O. A., Eriksson, K. S., Haas, H. L., and Brown, R. E. (2003). Excitation of ventral tegmental area dopaminergic and nondopaminergic neurons by orexins/hypocretins. *J Neurosci*, 23(1):7–11.

Korte, S. M., Koolhaas, J. M., Wingfield, J. C., and McEwen, B. S. (2005). The darwinian concept of stress: benefits of allostasis and costs of allostatic load and the trade-offs in health and disease. *Neuroscience & Biobehavioral Reviews*, 29(1):3–38.

Lapish, C., Kroener, S., Durstewitz, D., Lavin, A., and Seamans, J. (2007). The ability of the mesocortical dopamine system to operate in distinct temporal modes. *Psychopharmacology*, 191(3):609–625–625.

Lavin, A., Nogueira, L., Lapish, C. C., Wightman, R. M., Phillips, P. E., and Seamans, J. K. (2005). Mesocortical dopamine neurons operate in distinct temporal domains using multimodal signaling. *J Neurosci*, 25(20):5013–5023.

Lehrer, J. (2008). A new state of mind.

Leng, G. and Ludwig, M. (2006). Information processing in the hypothalamus: Peptides and analogue computation. *Journal of Neuroendocrinology*, 18(6):379–392.

Leshan, R. L., Opland, D. M., Louis, G. W., Leinninger, G. M., Patterson, C. M., Rhodes, C. J., Münzberg, H., and Myers, M. G. (2010). Ventral tegmental area leptin receptor neurons specifically project to and regulate cocaine- and amphetamine-regulated transcript neurons of the extended central amygdala. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30(16):5713–5723.

Lewis, D. A., Melchitzky, D. S., Sesack, S. R., Whitehead, R. E., Auh, S., and Sampson, A. (2001). Dopamine transporter immunoreactivity in monkey cerebral cortex: Regional, laminar, and ultrastructural localization. *J. Comp. Neurol.*, 432(1):119–136.

Lindskog, M., Kim, M., Wikström, M. A., Blackwell, K. T., and Kotaleski, J. H. (2006). Transient calcium and dopamine increase PKA activity and DARPP-32 phosphorylation. *PLoS Computational Biology*, 2(9):e119+.

Lockery, S. R., Goodman, M. B., and Faumont, S. (2009). First report of action potentials in a c. elegans neuron is premature. *Nature Neuroscience*, 12(4):365–366.

Lokhorst, G.-J. (2009). Descartes and the pineal gland. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Spring 2009 edition.

Lorenz, E. N. (1964). The problem of deducing the climate from the governing equations. *Tellus*, 16(1):1–11.

Ludvig, E. A., Sutton, R. S., and Kehoe, E. J. (2008). Stimulus representation and the timing of Reward-Prediction errors in models of the dopamine system. *Neural Computation*, 20(12):3034–3054.

Maier, S. (2003). Bi-directional immune-brain communication: Implications for understanding stress, pain, and cognition. *Brain, Behavior, and Immunity*, 17(2):69–85.

Maier, S. F. and Watkins, L. R. (1998). Cytokines for psychologists: implications of bidirectional immune-to-brain communication for understanding behavior, mood, and cognition. *Psychological review*, 105(1):83–107.

Marder, E. and Bucher, D. (2007). Understanding circuit dynamics using the stomatogastric nervous system of lobsters and crabs. *Annual Review of Physiology*, 69(1):291–316.

Margolis, E. B., Lock, H., Hjelmstad, G. O., and Fields, H. L. (2006). The ventral tegmental area revisited: is there an electrophysiological marker for dopaminergic neurons? *The Journal of physiology*, 577(Pt 3):907–924.

Marr, D. (1969). A theory of cerebellar cortex. *The Journal of Physiology*, 202(2):437–470.

Marr, D. (1983). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt & Company.

May, R. M. (1976). Simple mathematical models with very complicated dynamics. *Nature*, 261(5560):459–467.

McAfoose, J. and Baune, B. T. (2009). Evidence for a cytokine model of cognitive function. *Neuroscience and biobehavioral reviews*, 33(3):355–366.

McEwen, B. S. (1998). Stress, adaptation, and disease: Allostasis and allostatic load. *Annals of the New York Academy of Sciences*, 840(1):33–44.

Mehta, M. A., Manes, F. F., Magnolfi, G., Sahakian, B. J., and Robbins, T. W. (2004). Impaired set-shifting and dissociable effects on tests of spatial working memory following the dopamine d 2 receptor antagonist sulpiride in human volunteers. *Psychopharmacology*, 176(3):331–342.

Mesce, K. A. (2002). Metamodulation of the biogenic amines: second-order modulation by steroid hormones and amine cocktails. *Brain, behavior and evolution*, 60(6):339–349.

Montague, P. R., Dayan, P., and Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *J. Neurosci.*, 16(5):1936–1947.

Morimoto, M. (1996). Distribution of glucocorticoid receptor immunoreactivity and mRNA in the rat brain: an immunohistochemical and in situ hybridization study. *Neuroscience Research*, 26(3):235–269.

Mountcastle, V. B. (1957). Modality and topographic properties of single neurons of cat's somatic sensory cortex. *Journal of neurophysiology*, 20(4):408–434.

Nakamura, K., Matsumoto, M., and Hikosaka, O. (2008). Reward-Dependent modulation of neuronal activity in the primate dorsal raphe nucleus. *J. Neurosci.*, 28(20):5331–5343.

Nakano, T., Doi, T., Yoshimoto, J., and Doya, K. (2010). A kinetic model of dopamine- and calcium-dependent striatal synaptic plasticity. *PLoS computational biology*, 6(2).

Neve, K. A., Seamans, J. K., and Trantham-Davidson, H. (2004). Dopamine receptor signaling. *J Recept Signal Transduct Res*, 24(3):165–205.

O'Carroll, A.-M., Fowler, C. J., Phillips, J. P., Tobbia, I., and Tipton, K. F. (1983). The deamination of dopamine by human brain monoamine oxidase. *Naunyn-Schmiedeberg's Archives of Pharmacology*, 322(3):198–202–202.

Okasha, S. (2009). Biological altruism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2009 edition.

Olney, J. W., Newcomer, J. W., and Farber, N. B. (1999). NMDA receptor hypofunction model of schizophrenia. *Journal of psychiatric research*, 33(6):523–533.

O'Reilly, R. C., Norman, K. A., and McClelland, J. L. (1998). A hippocampal model of recognition memory. In *NIPS '97: Proceedings of the 1997 conference on Advances in neural information processing systems 10*, pages 73–79, Cambridge, MA, USA. MIT Press.

Otani, S., Blond, O., Desce, J. M., and Crépel, F. (1998). Dopamine facilitates long-term depression of glutamatergic transmission in rat prefrontal cortex. *Neuroscience*, 85(3):669–676.

Paspalas, C. D. and Goldman-Rakic, P. S. (2005). Presynaptic d1 dopamine receptors in primate prefrontal cortex: target-specific expression in the glutamatergic synapse. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 25(5):1260–1267.

Pawlak, V. and Kerr, J. N. (2008). Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(10):2435–2446.

Penfield, W. and Jasper, H. (1954). *Epilepsy and the Functional Anatomy of the Human Brain*. Little Brown & Co.

Perelson, A. S. and Weisbuch, G. (1997). Immunology for physicists. *Reviews of Modern Physics*, 69(4):1219+.

Petitto, J. (1997). Modulation of behavioral and neurochemical measures of forebrain dopamine function in mice by species-specific interleukin-2. *Journal of Neuroimmunology*, 73(1-2):183–190.

Peyron, C., Tighe, D. K., van den Pol, A. N., de Lecea, L., Heller, H. C., Sutcliffe, J. G., and Kilduff, T. S. (1998). Neurons containing hypocretin (orexin) project to multiple neuronal systems. *The Journal of Neuroscience*, 18(23):9996–10015.

Piazza, P. V., Rougé-Pont, F., Deroche, V., Maccari, S., Simon, H., and Le Moal, M. (1996). Glucocorticoids have state-dependent stimulant effects on the mesencephalic dopaminergic transmission. *Proceedings of the National Academy of Sciences of the United States of America*, 93(16):8716–8720.

Piccolino, M. (1998). Animal electricity and the birth of electrophysiology: the legacy of luigi galvani. *Brain Research Bulletin*, 46(5):381–407.

Pothos, E. N., Davila, V., and Sulzer, D. (1998). Presynaptic recording of quanta from midbrain dopamine neurons and modulation of the quantal size. *J. Neurosci.*, 18(11):4106–4118.

Pulvermüller, F. (1999). Words in the brain's language. *The Behavioral and brain sciences*, 22(2).

Quan, N. and Banks, W. (2007). Brain-immune communication pathways. *Brain, Behavior, and Immunity*, 21(6):727–735.

Rao, S. G., Williams, G. V., and Goldman-Rakic, P. S. (1999). Isodirectional tuning of adjacent interneurons and pyramidal cells during working memory: evidence for microcolumnar organization in PFC. *J Neurophysiol*, 81(4):1903–1916.

Redgrave, P., Gurney, K., and Reynolds, J. (2008). What is reinforced by phasic dopamine signals? *Brain research reviews*, 58(2):322–339.

Rescorla, R. A. and Wagner, A. W. (1972). *A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement*, chapter 3, pages 64–99. Appleton-Century-Crofts, New York.

Richfield, E. K., Penney, J. B., and Young, A. B. (1989). Anatomical and affinity state comparisons between dopamine d1 and d2 receptors in the rat central nervous system. *Neuroscience*, 30(3):767–777.

Rizzolatti, G. and Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27(1):169–192.

Robbins, T. W. (2005). Chemistry of the mind: neurochemical modulation of prefrontal cortical function. *J Comp Neurol*, 493(1):140–146.

Robinson, D. L., Hermans, A., Seipel, A. T., and Wightman, R. M. (2008). Monitoring rapid chemical communication in the brain. *Chemical Reviews*, 108(7):2554–2584.

Rosenblueth, A., Wiener, N., and Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of Science*, 10(1):18–24.

Rougé-Pont, F., Abrous, D. N., Le Moal, M., and Piazza, P. V. (1999). Release of endogenous dopamine in cultured mesencephalic neurons: influence of dopaminergic agonists and glucocorticoid antagonists. *The European journal of neuroscience*, 11(7):2343–2350.

Rougé-Pont, F., Deroche, V., Le Moal, M., and Piazza, P. V. (1998). Individual differences in stress-induced dopamine release in the nucleus accumbens are influenced by corticosterone. *The European journal of neuroscience*, 10(12):3903–3907.

Rubin, J. E., Gerkin, R. C., Bi, G. Q., and Chow, C. C. (2005). Calcium time course as a signal for spike-timing-dependent plasticity. *J Neurophysiol*, 93(5):2600–2613.

Savin, C. and Triesch, J. (2009). Developing a working memory with reward-modulated STDP.

Schildkraut, J. J. (1965). The catecholamine hypothesis of affective disorders: A review of supporting evidence. *Am J Psychiatry*, 122(5):509–522.

Schiller, J., Schiller, Y., and Clapham, D. E. (1998). NMDA receptors amplify calcium influx into dendritic spines during associative pre- and postsynaptic activation. *Nature Neuroscience*, 1(2):114–118.

Schneider, H., Pitossi, F., Balschun, D., Wagner, A., del Rey, A., and Besedovsky, H. O. (1998). A neuromodulatory role of interleukin-1beta in the hippocampus. *Proceedings of the National Academy of Sciences of the United States of America*, 95(13):7778–7783.

Schultz, W. (1986). Responses of midbrain dopamine neurons to behavioral trigger stimuli in the monkey. *J Neurophysiol*, 56(5):1439–1461.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80(1):1–27.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, 36(2):241–263.

Schultz, W. (2007). Behavioral dopamine signals. *Trends in Neurosciences*, 30(5):203–210.

Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *The Journal of Neuroscience*, 13(3):900–913.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science (New York, N.Y.)*, 275(5306):1593–1599.

Schummers, J., Yu, H., and Sur, M. (2008). Tuned responses of astrocytes and their influence on hemodynamic signals in the visual cortex. *Science*, 320(5883):1638–1643.

Seamans, J. (2007). Dopamine anatomy. *Scholarpedia*, 2(6):3737.

Seamans, J. and Durstewitz, D. (2008). Dopamine modulation. *Scholarpedia*, 3(4):2711.

Seamans, J. K., Durstewitz, D., Christie, B. R., Stevens, C. F., and Sejnowski, T. J. (2001). Dopamine D1/D5 receptor modulation of excitatory synaptic inputs to layer v prefrontal cortex neurons. *PNAS*, 98(1):301–306.

Seamans, J. K. and Yang, C. R. (2004). The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Progress in neurobiology*, 74(1):1–58.

Seeman, P. and Vantol, H. (1994). Dopamine receptor pharmacology. *Trends in Pharmacological Sciences*, 15(7):264–270.

Selye, H. (1955). Stress and Disease. *Science*, 122:625–631.

Selye, H. (1975). Confusion and controversy in the stress field. *Journal of human stress*, 1(2):37–44.

Sesack, S. R., Hawrylak, V. A., Matus, C., Guido, M. A., and Levey, A. I. (1998). Dopamine axon varicosities in the prelimbic division of the rat prefrontal cortex exhibit sparse immunoreactivity for the dopamine transporter. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 18(7):2697–2708.

Seutin, V. (2005). Dopaminergic neurones: much more than dopamine? *Br J Pharmacol*, 146(2):167–169.

Shen, W., Flajolet, M., Greengard, P., and Surmeier, D. J. (2008). Dichotomous dopaminergic control of striatal synaptic plasticity. *Science (New York, N.Y.)*, 321(5890):848–851.

Sherman, S. M. and Guillery, R. W. (1998). On the actions that one nerve cell can have on another: Distinguishing drivers from modulators. *Proceedings of the National Academy of Sciences of the United States of America*, 95(12):7121–7126.

Shouval, H. Z. (2007). Models of synaptic plasticity. *Scholarpedia*, 2(7):1605.

Shouval, H. Z., Bear, M. F., and Cooper, L. N. (2002). A unified model of NMDA receptor-dependent bidirectional synaptic plasticity. *Proceedings of the National Academy of Sciences of the United States of America*, 99(16):10831–10836.

Shuler, M. G. and Bear, M. F. (2006). Reward timing in the primary visual cortex. *Science*, 311(5767):1606–1609.

Sjöström, P. J., Turrigiano, G. G., and Nelson, S. B. (2001). Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron*, 32(6):1149–1164.

Skinner, B. F. (1976). *About Behaviorism*. Vintage, 1 edition.

Smiley, J. F., Levey, A. I., Ciliax, B. J., and Goldman-Rakic, P. S. (1994). D1 dopamine receptor immunoreactivity in human and monkey cerebral cortex: predominant and extrasynaptic localization in dendritic spines. *Proceedings of the National Academy of Sciences of the United States of America*, 91(12):5720–5724.

Smolensky, P. and Legendre, G. (2006). *The Harmonic Mind: From Neural Computation to Optimality-Theoretic GrammarVolume I: Cognitive Architecture (Bradford Books)*. The MIT Press.

Song, C. (1999). Variations of nucleus accumbens dopamine and serotonin following systemic interleukin-1, interleukin-2 or interleukin-6 treatment. *Neuroscience*, 88(3):823–836.

Spielewoy, C., Roubert, C., Hamon, M., Nosten-Bertrand, M., Betancur, C., and Giros, B. (2000). Behavioural disturbances associated with hyperdopaminergia

in dopamine-transporter knockout mice. *Behavioural pharmacology*, 11(3-4):279–290.

Spruston, N., Schiller, Y., Stuart, G., and Sakmann, B. (1995). Activity-dependent action potential invasion and calcium influx into hippocampal CA1 dendrites. *Science*, 268(5208):297–300.

Stellwagen, D. and Malenka, R. C. (2006). Synaptic scaling mediated by glial TNF-alpha. *Nature*, 440(7087):1054–1059.

Suri, R. (2002). TD models of reward predictive responses in dopamine neurons. *Neural Networks*, 15(4-6):523–533.

Suri, R. E. and Schultz, W. (1998). Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Experimental Brain Research*, 121(3):350–354.

Surmeier, D. J., Song, W. J., and Yan, Z. (1996). Coordinated expression of dopamine receptors in neostriatal medium spiny neurons. *J Neurosci*, 16(20):6579–6591.

Svenningsson, P., Nishi, A., Fisone, G., Girault, J.-A. A., Nairn, A. C., and Greengard, P. (2004). DARPP-32: an integrator of neurotransmission. *Annual review of pharmacology and toxicology*, 44:269–296.

Szelényi, J. (2001). Cytokines and the central nervous system. *Brain Research Bulletin*, 54(4):329–338.

Tauber, A. (2000). Moving beyond the immune self? *Seminars in Immunology*, 12(3):241–248.

Tauber, A. (2010). The biological notion of self and non-self. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Summer 2010 edition.

Tegnér, J., Compte, A., and Wang, X. J. (2002). The dynamical stability of reverberatory neural circuits. *Biol Cybern*, 87(5-6):471–481.

Tepper, J. M. and Plenz, D. (2006). *Microcircuits in the striatum: Striatal cell types and their interaction*. The MIT Press.

Thivierge, J.-P., Rivest, F., and Monchi, O. (2007). Spiking neurons, dopamine, and plasticity: Timing is everything, but concentration also matters. *Synapse*, 61(6):375–390.

Thurley, K., Senn, W., and Luscher, H.-R. (2008). Dopamine increases the gain of the Input-Output response of rat prefrontal pyramidal neurons. *J Neurophysiol*, 99(6):2985–2997.

Tobler, P. N., Fiorillo, C. D., and Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science*, 307(5715):1642–1645.

Tseng, K. Y., Mallet, N., Toreson, K. L., Le Moine, C., Gonon, F., and O'Donnell, P. (2006). Excitatory response of prefrontal cortical fast-spiking interneurons to ventral tegmental area stimulation in vivo. *Synapse*, 59(7):412–417.

Varela, F. G., Maturana, H. R., and Uribe, R. (1974). Autopoiesis: the organization of living systems, its characterization and a model. *Currents in modern biology*, 5(4):187–196.

Vizi, S. E., Kiss, J. P., and Lendvai, B. (2004). Nonsynaptic communication in the central nervous system. *Neurochemistry International*, 45(4):443–451.

Voorn, P., Vanderschuren, L. J., Groenewegen, H. J., Robbins, T. W., and Pennartz, C. M. (2004). Putting a spin on the dorsal-ventral divide of the striatum. *Trends Neurosci*, 27(8):468–474.

Wallis, J. D. and Miller, E. K. (2003). Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *European Journal of Neuroscience*, 18(7):2069–2081.

Wang, B., Shaham, Y., Zitzman, D., Azari, S., Wise, R. A., and You, Z.-B. (2005). Cocaine experience establishes control of midbrain glutamate and dopamine by Corticotropin-Releasing factor: A role in Stress-Induced relapse to drug seeking. *The Journal of Neuroscience*, 25(22):5389–5396.

Wang, B., You, Z.-B., Rice, K., and Wise, R. (2007). Stress-induced relapse to cocaine seeking: roles for the CRF₂ receptor and CRF-binding protein in the ventral tegmental area of the rat. *Psychopharmacology*, 193(2):283–294.

Wang, M., Vijayraghavan, S., and Goldman-Rakic, P. S. (2004). Selective d2 receptor actions on the functional circuitry of working memory. *Science*, 303(5659):853–856.

Westerink, B. H., Kwint, H. F., and deVries, J. B. (1996). The pharmacology of mesolimbic dopamine neurons: a dual-probe microdialysis study in the ventral tegmental area and nucleus accumbens of the rat brain. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 16(8):2605–2611.

White, F. J. (1996). Synaptic regulation of mesocorticolimbic dopamine neurons. *Annu Rev Neurosci*, 19:405–436.

Willie, J. T., Chemelli, R. M., Sinton, C. M., and Yanagisawa, M. (2001). To eat or to sleep? orexin in the regulation of feeding and wakefulness. *Annual review of neuroscience*, 24(1):429–458.

Winslow, J. T., Hastings, N., Carter, C. S., Harbaugh, C. R., and Insel, T. R. (1993). A role for central vasopressin in pair bonding in monogamous prairie voles. *Nature*, 365(6446):545–548.

Xu, T.-X. and Yao, W.-D. (2010). D1 and d2 dopamine receptors in separate circuits cooperate to drive associative long-term potentiation in the prefrontal cortex. *Proceedings of the National Academy of Sciences*, 107(37):16366–16371.

Ye, J., Zalcman, S., and Tao, L. (2005). Kainate-activated currents in the ventral tegmental area of neonatal rats are modulated by interleukin-2. *Brain Research*, 1049(2):227–233.

Yu, A. and Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, 46(4):681–692.